Text Visualization for the Support of Lexicography-Based Scholarly Work

Shane Sheehan, Saturnino Luz

Usher Institute of Population Health Sciences & Informatics,
The University of Edinburgh, UK
E-mail: Shane.Sheehan@ed.ac.uk, S.luz@ed.ac.uk

Abstract

We discuss three visualisation techniques for corpus analysis, Concordance Mosaic, Metafacet and ComFre, and explore the design rationale based on a characterization of the corpus linguistic domain. The Concordance Mosaic visualization is designed for the investigation of collocation patterns. It encodes word positions in a concordance list in a manner that emphasizes quantitative analysis of frequency or collocation statistics. Metafacet provides an interface for investigating concordance lists through the lens of meta-data. When combined with the Mosaic it provides a powerful technique for investigating collocations in the context of meta-data. ComFre can be used to compare word frequencies between two corpora of different size, it has potential use as a technique for identifying terms which are representative of the corpora under investigation. The domain characterization shows how the visualizations were designed with corpus linguistic methodologies at the core. It consists of a task analysis based on the methodology outlined in Sinclairs' *Reading Concordances: An Introduction*, and the analysis of methodology case studies from language scholars.

Keywords: visualization; concordance; frequency; meta-data; collocation

1. Introduction

Concordance analysis is a core activity of scholars in a number of humanities disciplines, including corpus linguistics, classical studies, and translation studies, to name a few. Through the advent of technology and the ever increasing availability of textual data this type of structured analysis of text has grown in importance (Sinclair, 1991; Bonelli, 2010).

In concordance analysis, every corpus occurrence of a keyword of interest is displayed along with its context. The context is an ordered list of words which precede and follow the keyword. The analyst then seeks to discover the linguistic properties of the keyword and the contextual patterns which predict them by observing the frequencies of occurrence, in the keyword's context, of words (collocations), word combinations, parts of speech (colligations) or the various other lexical classifications (Sinclair, 2003; Scott, 2010).

The most widely used tool in this kind of analysis is a form of tabular visualization known as keyword-in-context (KWIC). The creation of concordances through the keyword in context indexing technique was first proposed by Hans Peter Luhn in the 1950's (Luhn & Division, 1959). KWIC displays, enhanced in interactive systems by features such as search, context sorting and statistical analysis, are widely used not only by academics and scholars, but also by professional translators and post-editors (Karamanis et al., 2011; Doherty et al., 2012).

While these KWIC interfaces provide support for exploring the linear structure of the concordance, word frequency and other statistics rarely form any part of the visualization. This statistical information is essential to the work of the text analyst. However, in the presence of large corpora, it is difficult to explore statistical regularities armed solely with the KWIC display. External statistical tools are often used to complement the concordance. We argue that integration of this analysis step into the concordance visualization fits in well with the task structure of corpus linguists, and will be of great benefit to the text analyst.

There have been calls for the creation of more advanced concordance analysis tools (Rockwell, 2003), and advancements such as Sketch Engine have provide new analytic paths (Kilgarriff et al., 2014). However, the adoption of visual analysis tools for concordance analysis is very limited. That does not mean that visual representations of the concordance do not exist, it is simply that they have not been adopted by analysts or integrated into analysis tools.

It has been suggested that the publication of more domain characterization papers for visualization would be beneficial for tool adoption (Munzner, 2009). It is at this level of design that relevant problems are identified, and creating visual solutions to problems that are not relevant to domain experts is wasted effort. Publication of domain characterization should also encourage wider conversation and help identify and characterize overlooked areas of investigation.

In this paper we outline the functionality of three corpus analysis tools, Concordance Mosaic, Metafacet and ComFre. Concordance Mosaic displays positional collocation statistics for any corpus word or regular expression. Interactive restructuring of a concordance browser is enabled through the interface. This restructuring combined with colour highlighting of the concordance lines creates a powerful technique for investigating significant collocation patterns.

The MetaFacet visualization enables exploration of corpora through the lens of meta-data. Keyword frequency can be investigated across any combination of meta-data attributes associated with corpus source files. The concordance browser and Mosaic can be interactively filtered by these attribute combinations, allowing investigation and comparison of lexical information across combinations such as date, author and topic.

ComFre is a tool for corpus frequency comparison, which provides a method of comparing corpora of different size in a visual and statistically valid manner.

These visualizations were designed in close collaboration with language scholars with an emphasis on translation studies. The design rationale is rooted in a domain characterization which encompasses a literature-based task analysis and ethnographic studies of methodology. Relevant portions of this domain characterization are presented following the visualization descriptions.

2. Modnlp plugins

The visualization tools are developed as plugins for the open source concordance browser included in the Modnlp toolkit. Significant contributions were also made to the core Modnlp project to better integrate the plugins and enable interactions with the concordance list. Modnlp provides a modular architecture and tools for natural language processing, it comes with an indexer, feature rich concordance browser and server implementation (Luz, 2011, 2000). Previous versions of the Modnlp software have been used by the European Parliamentary Comparable and Parallel Corpora project[1] (ECPC) and by the Translational English Corpus[2] (TEC). The toolkit is currently being developed as part of the Genealogies of Knowledge project[3] (GoK) and the plugins are fully integrated into the GoK corpus browser.

The goal when developing these plugins is to improve the efficiency and capability of corpus linguistic methodologies and tools. Here we present the visualization plugins from a purely functional perspective to provide an overview of the capabilities and context for later discussion of the relevance to lexicography and corpus linguistics.

The English GoK corpus is used to exemplify the usage of the visualizations. This corpus is quite varied, it includes translations and re-translations of texts from antiquity as well as modern internet blogs and magazine articles. The corpus is designed to enable researchers to trace the trajectory of key concepts as they enter different cultural and temporal spaces, predominantly but not exclusively through the mediation of various forms of translation. The corpus is specialized and the examples used may not exhibit general lexical properties due to the issues of representativeness in relation to frequency (Summers, 1996).

In the discussion of the visualization functionality we do not try to analyse or interpret the linguistic properties of the words or corpus. Any analysis choice or comments on linguistic properties are to help clarify the examples and should not be viewed as an attempt to perform corpus analysis.

[1] http://www.ecpc.uji.es/

[2] http://genealogiesofknowledge.net/translational-english-corpus-tec/

[3] http://genealogiesofknowledge.net/

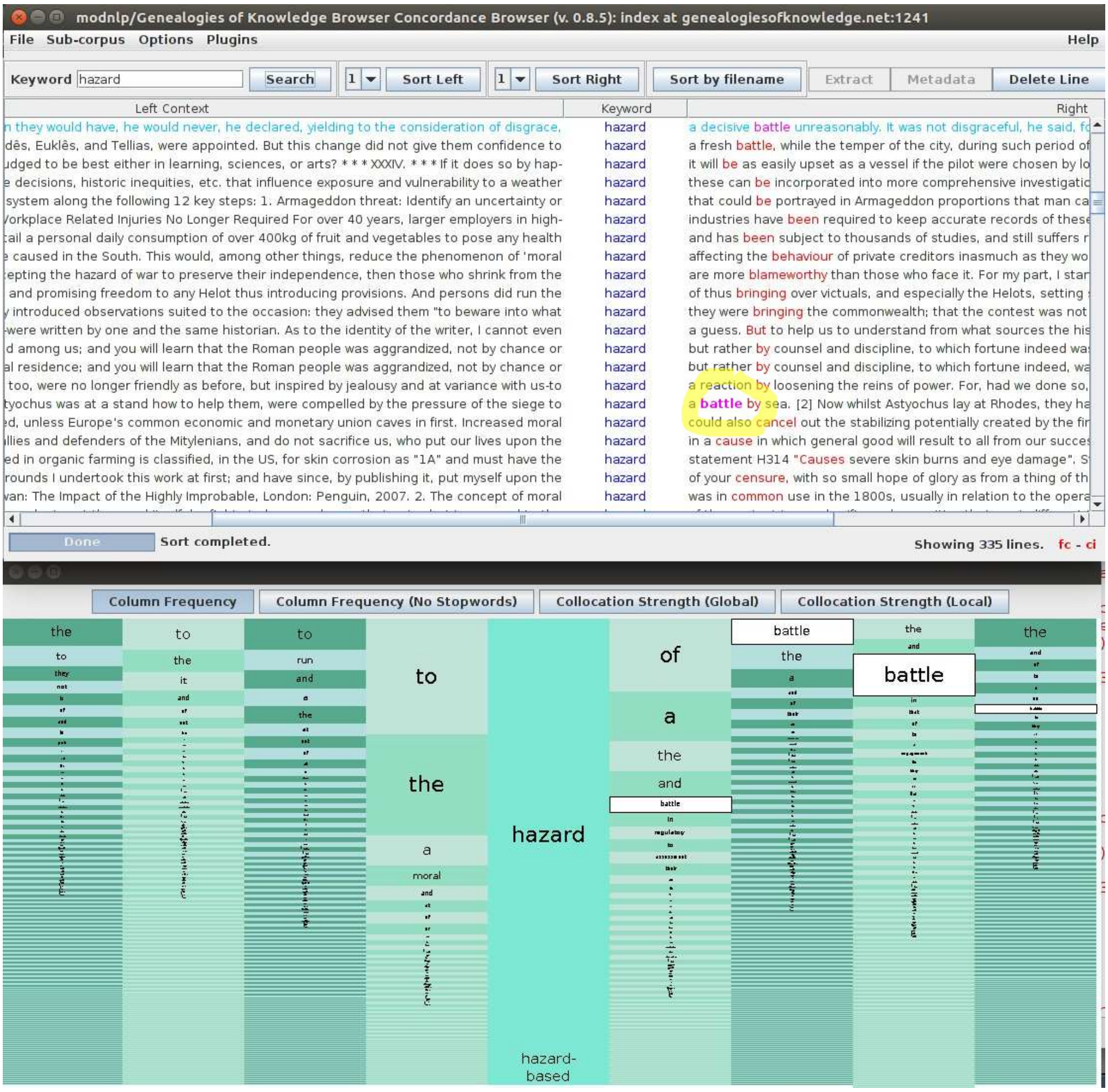
2.1 Concordance Mosaic

The first visualization designed was the Concordance Mosaic. This visualization has the concept of keyword in context at its core. The visualization is designed to display word statistics per position extracted from a concordance list. The underlying graph based abstraction of the concordance list and an early prototype were presented in an earlier work (Luz & Sheehan, 2014).

Using the visual metaphor of the KWIC, Mosaic represents positions relative to the keyword as ordered columns of tiles. The mosaic is created using a space-filling approach introduced by Luz and Masoodian (2007), where each tile represents a word at a position relative to the keyword, and the height of each tile is proportional to the word statistic at that position. In its simplest form each tile represents the frequency of a word at a position relative to the keyword. In Figure 1 the Mosaic of the keyword “hazard” is presented along with the concordance list for the 335 occurrences in the corpus. The Mosaic is set to display column frequencies. Due to the strong visual metaphor of KWIC it should be clear the word “to” is the most frequent word immediately to the left of the keyword (K-1) and also at positions K-2 and K-3. Hovering over any tile will display a tool-tip with the word count and frequency at the position, this relieves the need for manually counting or performing additional searches to retrieve position based word frequencies.

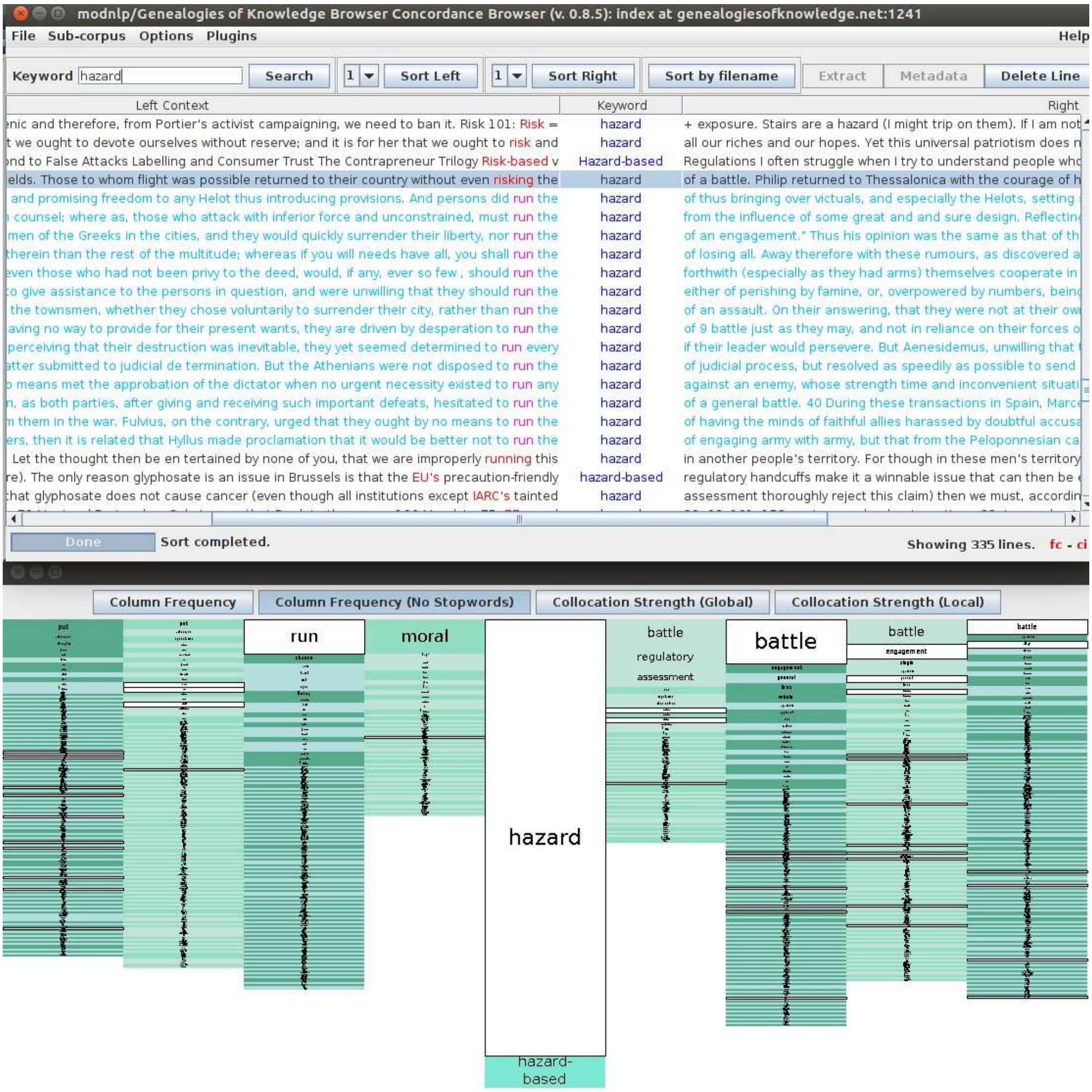
Words with high corpus frequency tend to dominate the positional frequency distributions for most keywords. The second view Mosaic affords is a stop-word filtered view of column frequency. The columns are filtered using a threshold based on corpus frequency. In Figure 2, the stop-words are removed and column heights are no longer uniform. The reduction in a column’s height represents the density of stop-word frequency at that position. At K-1 we notice stop-words were the most frequent for any position. At K-1 the next most frequent word after stop-words is “moral”. Tile heights and thus frequency are comparable across positions, from the Mosaic we can see that “moral” at position K-1 and “run” at position K-2 have similar positional frequencies.

The mosaic and concordance browser have been presented together but we have not yet commented on the interaction. The data is linked to both interfaces, and interactions with the mosaic can be reflected on the concordance list. In Figure 2 the tile for the word “run” at K-2 has been left clicked with the mouse. This interaction colours white any position word tiles on the Mosaic that are found in concordance lines, including “run” at K-2.



Left Context	Keyword	Right
n they would have, he would never, he declared, yielding to the consideration of disgrace,	hazard	a decisive battle unreasonably. It was not disgraceful, he said, fo
dês, Euklês, and Tellias, were appointed. But this change did not give them confidence to	hazard	a fresh battle, while the temper of the city, during such period of
udged to be best either in learning, sciences, or arts? * * * XXXIV. * * * If it does so by hap-	hazard	it will be as easily upset as a vessel if the pilot were chosen by lo
e decisions, historic inequities, etc. that influence exposure and vulnerability to a weather	hazard	these can be incorporated into more comprehensive investigatio
system along the following 12 key steps: 1. Armageddon threat: Identify an uncertainty or	hazard	that could be portrayed in Armageddon proportions that man ca
Vorkplace Related Injuries No Longer Required For over 40 years, larger employers in high-	hazard	industries have been required to keep accurate records of these
tail a personal daily consumption of over 400kg of fruit and vegetables to pose any health	hazard	and has been subject to thousands of studies, and still suffers r
e caused in the South. This would, among other things, reduce the phenomenon of 'moral	hazard	affecting the behaviour of private creditors inasmuch as they wo
epting the hazard of war to preserve their independence, then those who shrink from the	hazard	are more blameworthy than those who face it. For my part, I star
and promising freedom to any Helot thus introducing provisions. And persons did run the	hazard	of thus bringing over victuals, and especially the Helots, setting
y introduced observations suited to the occasion: they advised them "to beware into what	hazard	they were bringing the commonwealth; that the contest was not
-were written by one and the same historian. As to the identity of the writer, I cannot even	hazard	a guess. But to help us to understand from what sources the his
d among us; and you will learn that the Roman people was aggrandized, not by chance or	hazard	but rather by counsel and discipline, to which fortune indeed wa
al residence; and you will learn that the Roman people was aggrandized, not by chance or	hazard	but rather by counsel and discipline, to which fortune indeed, wa
too, were no longer friendly as before, but inspired by jealousy and at variance with us-to	hazard	a reaction by loosening the reins of power. For, had we done so,
tyochus was at a stand how to help them, were compelled by the pressure of the siege to	hazard	a battle by sea. [2] Now whilst Astyochus lay at Rhodes, they ha
ed, unless Europe's common economic and monetary union caves in first. Increased moral	hazard	could also cancel out the stabilizing potentially created by the fir
llies and defenders of the Mitylenians, and do not sacrifice us, who put our lives upon the	hazard	in a cause in which general good will result to all from our succes
ed in organic farming is classified, in the US, for skin corrosion as "1A" and must have the	hazard	statement H314 "Causes severe skin burns and eye damage". S
rounds I undertook this work at first; and have since, by publishing it, put myself upon the	hazard	of your censure, with so small hope of glory as from a thing of th
van: The Impact of the Highly Improbable, London: Penguin, 2007. 2. The concept of moral	hazard	was in common use in the 1800s, usually in relation to the opera

Figure 1: Concordance Mosaic for keyword "hazard". Right click selection of "battle" at position K+3.



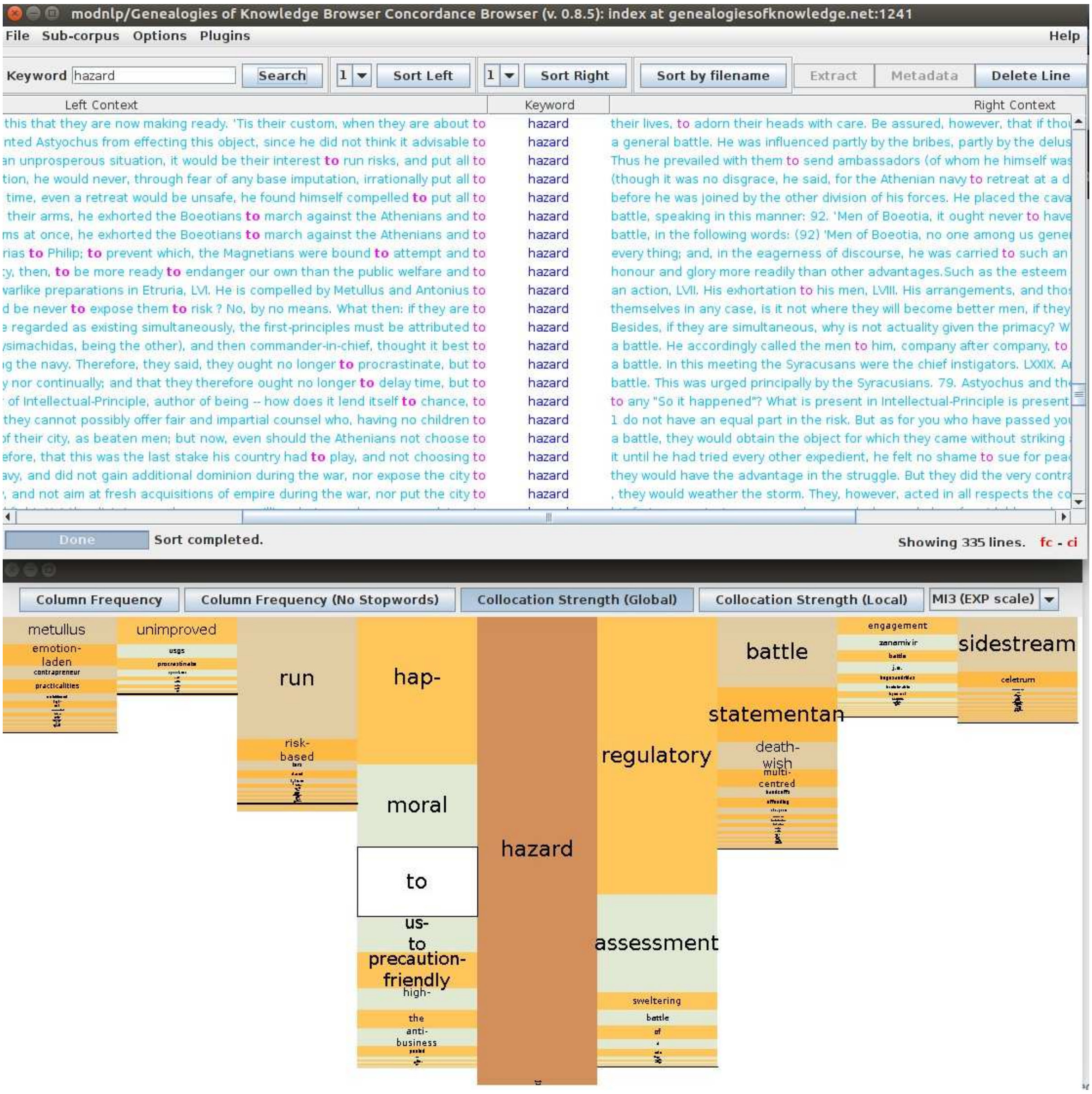
Left Context	Keyword	Right
ɛnic and therefore, from Portier's activist campaigning, we need to ban it. Risk 101: Risk =	hazard	+ exposure. Stairs are a hazard (I might trip on them). If I am not
t we ought to devote ourselves without reserve; and it is for her that we ought to risk and	hazard	all our riches and our hopes. Yet this universal patriotism does n
ɔnd to False Attacks Labelling and Consumer Trust The Contrapreneur Trilogy Risk-based v	Hazard-based	Regulations I often struggle when I try to understand people who
elds. Those to whom flight was possible returned to their country without even risking the	hazard	of a battle. Philip returned to Thessalonica with the courage of h
and promising freedom to any Helot thus introducing provisions. And persons did run the	hazard	of thus bringing over victuals, and especially the Helots, setting
counsel; where as, those who attack with inferior force and unconstrained, must run the	hazard	from the influence of some great and and sure design. Reflecting
men of the Greeks in the cities, and they would quickly surrender their liberty, nor run the	hazard	of an engagement." Thus his opinion was the same as that of th
therein than the rest of the multitude; whereas if you will needs have all, you shall run the	hazard	of losing all. Away therefore with these rumours, as discovered a
even those who had not been privy to the deed, would, if any, ever so few , should run the	hazard	forthwith (especially as they had arms) themselves cooperate in
to give assistance to the persons in question, and were unwilling that they should run the	hazard	either of perishing by famine, or, overpowered by numbers, being
the townsmen, whether they chose voluntarily to surrender their city, rather than run the	hazard	of an assault. On their answering, that they were not at their ow
aving no way to provide for their present wants, they are driven by desperation to run the	hazard	of 9 battle just as they may, and not in reliance on their forces o
perceiving that their destruction was inevitable, they yet seemed determined to run every	hazard	if their leader would persevere. But Aenesidemus, unwilling that
atter submitted to judicial de termination. But the Athenians were not disposed to run the	hazard	of judicial process, but resolved as speedily as possible to send
o means met the approbation of the dictator when no urgent necessity existed to run any	hazard	against an enemy, whose strength time and inconvenient situati
n, as both parties, after giving and receiving such important defeats, hesitated to run the	hazard	of a general battle. 40 During these transactions in Spain, Marc
n them in the war. Fulvius, on the contrary, urged that they ought by no means to run the	hazard	of having the minds of faithful allies harassed by doubtful accusa
ers, then it is related that Hyllus made proclamation that it would be better not to run the	hazard	of engaging army with army, but that from the Peloponnesian ca
Let the thought then be en tertained by none of you, that we are improperly running this	hazard	in another people's territory. For though in these men's territory
re). The only reason glyphosate is an issue in Brussels is that the EU's precaution-friendly	hazard-based	regulatory handcuffs make it a winnable issue that can then be
that glyphosate does not cause cancer (even though all institutions except IARC's tainted	hazard	assessment thoroughly reject this claim) then we must, accordin

Sort completed.

Showing 335 lines.

Figure 2: Concordance Mosaic for keyword "hazard", stop words have been removed. Left click selection of "run" at position K-2.

Looking at the Mosaic we see that at least one concordance line with "run" at position K-2 also contains "battle" at K+2. The concordance list has been sorted at the selected position and scrolled automatically to the selected word. For emphasis the sorted position words are coloured red and the selected word coloured pink. The horizontal concordance lines for the selected word are coloured blue for easy identification. In addition, any occurrences of the selected word at other positions are also highlighted in pink, and as you investigate the entire list it is possible to get a sense of global patterns which may not be restricted to the selected position. In Figure 3 the selection of the word "to" at K-1 and a sample of its many occurrences at other positions are visible.


modnlp/Genealogies of Knowledge Browser Concordance Browser (v. 0.8.5): index at genealogiesofknowledge.net:1241
File Sub-corpus Options Plugins Help
Keyword hazard Search 1 Sort Left 1 Sort Right Sort by filename Extract Metadata Delete Line
Done Sort completed.
Showing 335 lines. fc - ci
Column Frequency
Column Frequency (No Stopwords)
Collocation Strength (Global)
Collocation Strength (Local)
MI3 (EXP scale)
metullus
emotion-laden
contrapreneur
practicalities
unimproved
usgs
procrastinate
run
risk-based
hap-
moral
to
us-to
precaution-friendly
high-
the
anti-business
hazard
regulatory
assessment
sweltering
battle
of
battle
statementan
death-wish
multi-centred
engagement
zanamivir
battle
j.e.
sidestream
celetrum

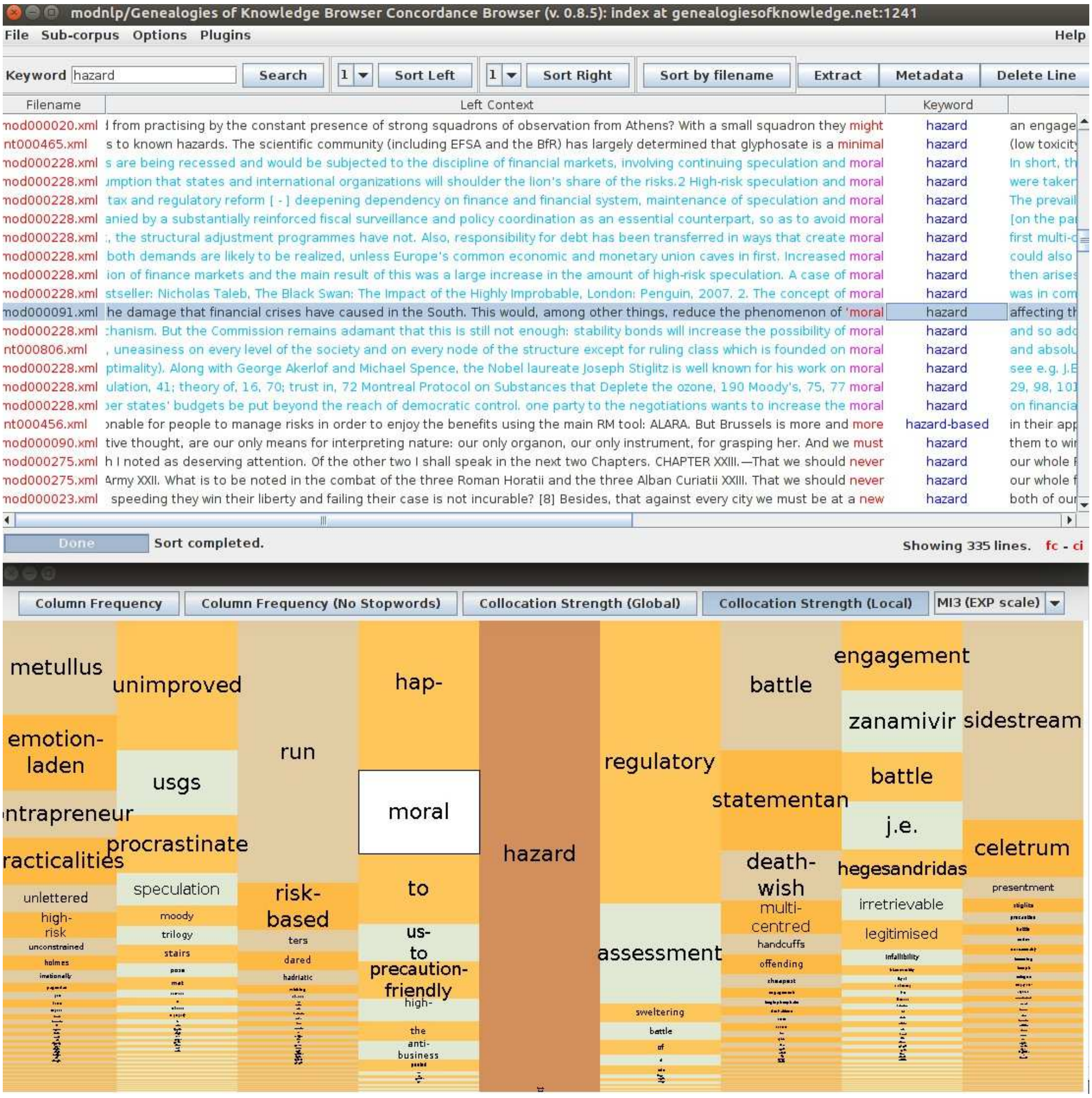
Left Context	Keyword	Right Context
this that they are now making ready. 'Tis their custom, when they are about to	hazard	their lives, to adorn their heads with care. Be assured, however, that if thou
nted Astyochus from effecting this object, since he did not think it advisable to	hazard	a general battle. He was influenced partly by the bribes, partly by the delus
an unprosperous situation, it would be their interest to run risks, and put all to	hazard	Thus he prevailed with them to send ambassadors (of whom he himself was
tion, he would never, through fear of any base imputation, irrationally put all to	hazard	(though it was no disgrace, he said, for the Athenian navy to retreat at a d
time, even a retreat would be unsafe, he found himself compelled to put all to	hazard	before he was joined by the other division of his forces. He placed the cava
their arms, he exhorted the Boeotians to march against the Athenians and to	hazard	battle, speaking in this manner: 92. 'Men of Boeotia, it ought never to have
ms at once, he exhorted the Boeotians to march against the Athenians and to	hazard	battle, in the following words: (92) 'Men of Boeotia, no one among us gene
rias to Philip; to prevent which, the Magnetians were bound to attempt and to	hazard	every thing; and, in the eagerness of discourse, he was carried to such an
:y, then, to be more ready to endanger our own than the public welfare and to	hazard	honour and glory more readily than other advantages.Such as the esteem
varlike preparations in Etruria, LVI. He is compelled by Metullus and Antonius to	hazard	an action, LVII. His exhortation to his men, LVIII. His arrangements, and thos
d be never to expose them to risk ? No, by no means. What then: if they are to	hazard	themselves in any case, is it not where they will become better men, if they
e regarded as existing simultaneously, the first-principles must be attributed to	hazard	Besides, if they are simultaneous, why is not actuality given the primacy? W
/simachidas, being the other), and then commander-in-chief, thought it best to	hazard	a battle. He accordingly called the men to him, company after company, to
g the navy. Therefore, they said, they ought no longer to procrastinate, but to	hazard	a battle. In this meeting the Syracusans were the chief instigators. LXXIX. A
y nor continually; and that they therefore ought no longer to delay time, but to	hazard	battle. This was urged principally by the Syracusians. 79. Astyochus and th
of Intellectual-Principle, author of being -- how does it lend itself to chance, to	hazard	to any "So it happened"? What is present in Intellectual-Principle is present
they cannot possibly offer fair and impartial counsel who, having no children to	hazard	1 do not have an equal part in the risk. But as for you who have passed you
of their city, as beaten men; but now, even should the Athenians not choose to	hazard	a battle, they would obtain the object for which they came without striking
efore, that this was the last stake his country had to play, and not choosing to	hazard	it until he had tried every other expedient, he felt no shame to sue for pea
avy, and did not gain additional dominion during the war, nor expose the city to	hazard	they would have the advantage in the struggle. But they did the very contra
, and not aim at fresh acquisitions of empire during the war, nor put the city to	hazard	, they would weather the storm. They, however, acted in all respects the co

Figure 3: Concordance Mosaic for keyword "hazard". Global view of MI3 is selected. Right click selection of "to" at position K-1.

The second click is activated by right clicking on a Mosaic tile. This interaction has the same effect on the concordance list as the left click interaction but differs in its change to the Mosaic. Right clicking on a Mosaic tile highlights other occurrences of the word at all positions in the mosaic. This is useful for getting a better sense of the frequency distribution of a word across all positions in a concordance list. In Figure 1, "battle" at K+3 is selected. Tiles representing "battle" at positions K+1 K+2 and K+4 are coloured white for easy identification. In the concordance list we can see one of these additional occurrences of "battle" at K+2 highlighted in pink.

Positional word frequency is a fundamental property of the concordance list, but other quantitative measures are used extensively to reason about collocations. Statistics such as Mutual Information (MI), Cubic Mutual Information (MI3) and Z-Score are often used to investigate collocation statistics in a window surrounding a keyword (Manning

& Schütze, 1999). This windowed approach most often groups word positions together and presents the results as a list. However we wish to preserve the positional aspect of these statistics and present them as a Mosaic. Figure 3 shows the *global collocation strength* view of Mosaic. Global in this setting means the tiles can be compared across positions and have not been scaled to fill the space. This contrasts with Figure 4 where the *local* view of collocation strength makes each column full height, and this allows easier investigation of each position but removes the ability to compare tiles across positions.



modnlp/Genealogies of Knowledge Browser Concordance Browser (v. 0.8.5): index at genealogiesofknowledge.net:1241

Keyword hazard 1 1

Filename	Left Context	Keyword	
nod000020.xml	l from practising by the constant presence of strong squadrons of observation from Athens? With a small squadron they might	hazard	an engage
nt000465.xml	s to known hazards. The scientific community (including EFSA and the BfR) has largely determined that glyphosate is a minimal	hazard	(low toxicity
nod000228.xml	s are being recessed and would be subjected to the discipline of financial markets, involving continuing speculation and moral	hazard	In short, th
nod000228.xml	ımption that states and international organizations will shoulder the lion's share of the risks.2 High-risk speculation and moral	hazard	were taker
nod000228.xml	tax and regulatory reform [-] deepening dependency on finance and financial system, maintenance of speculation and moral	hazard	The prevail
nod000228.xml	anied by a substantially reinforced fiscal surveillance and policy coordination as an essential counterpart, so as to avoid moral	hazard	[on the pa
nod000228.xml	:, the structural adjustment programmes have not. Also, responsibility for debt has been transferred in ways that create moral	hazard	first multi-c
nod000228.xml	both demands are likely to be realized, unless Europe's common economic and monetary union caves in first. Increased moral	hazard	could also
nod000228.xml	ion of finance markets and the main result of this was a large increase in the amount of high-risk speculation. A case of moral	hazard	then arises
nod000228.xml	stseller: Nicholas Taleb, The Black Swan: The Impact of the Highly Improbable, London: Penguin, 2007. 2. The concept of moral	hazard	was in com
nod000091.xml	he damage that financial crises have caused in the South. This would, among other things, reduce the phenomenon of 'moral	hazard	affecting th
nod000228.xml	:hanism. But the Commission remains adamant that this is still not enough: stability bonds will increase the possibility of moral	hazard	and so add
nt000806.xml	, uneasiness on every level of the society and on every node of the structure except for ruling class which is founded on moral	hazard	and absolu
nod000228.xml	ptimality). Along with George Akerlof and Michael Spence, the Nobel laureate Joseph Stiglitz is well known for his work on moral	hazard	see e.g. J.E
nod000228.xml	ulation, 41; theory of, 16, 70; trust in, 72 Montreal Protocol on Substances that Deplete the ozone, 190 Moody's, 75, 77 moral	hazard	29, 98, 101
nod000228.xml	ɔer states' budgets be put beyond the reach of democratic control. one party to the negotiations wants to increase the moral	hazard	on financia
nt000456.xml	ɔnable for people to manage risks in order to enjoy the benefits using the main RM tool: ALARA. But Brussels is more and more	hazard-based	in their app
nod000090.xml	tive thought, are our only means for interpreting nature: our only organon, our only instrument, for grasping her. And we must	hazard	them to wi
nod000275.xml	h I noted as deserving attention. Of the other two I shall speak in the next two Chapters. CHAPTER XXIII.—That we should never	hazard	our whole f
nod000275.xml	Army XXII. What is to be noted in the combat of the three Roman Horatii and the three Alban Curiatii XXIII. That we should never	hazard	our whole f
nod000023.xml	speeding they win their liberty and failing their case is not incurable? [8] Besides, that against every city we must be at a new	hazard	both of our

Done Sort completed.

Showing 335 lines. fc - ci

MI3 (EXP scale)

Figure 4: Concordance Mosaic for keyword "hazard". Local view of MI3 is selected. Right click selection of "moral" at position K-1. Concordance list scrolled horizontally to reveal filenames.

In the *Global* view shown in Figure 3 the column heights give an indication of the word positions relative to the keyword where the statistical association is highest. Each individual tile's height is proportional to the value of the statistic calculated for that word at that position. In this example MI3 is selected as the statistic under investigation. The strongest association based on MI3 is the word “regulatory” at K+1. It may be worth noting that the stop-word “to” is shown to have a strong association at K-1.

If we investigate the concordance lines of the tile “moral” at K-1 (since it has both high frequency and MI3 score) we find that all but two of its 14 occurrences originate from the same file, see Figure 4.

2.2 Metafacet

The Modnlp concordance browser presents the file-names along with concordance lines. An interaction is available in the browser to view meta-data about each file and section on a line by line basis. However, this is a time consuming and challenging process for the corpus analyst if the meta-data of a large number of lines need to be investigated. The Metfacet plugin is a proposed solution to this issue and provides interactive filtering of the concordance list and the Mosaic using all available meta-data facets.

The Metafacet interface is quite simple, and uses a horizontal bar chart to display concordance line frequency per meta-data attribute. An attribute is a possible value that a meta-data facet can take. As an example “Plato” is an Attribute of the Facet “author”. A drop-down list is used to choose which facet is displayed and the bars are sortable by frequency or lexicographical order and the window can be filtered using a sliding scale to view a smaller portion of the attributes. This conforms to the common visualization design practice of first presenting an overview, and then more detail on demand (Shneiderman, 1996).

In Figures 5 and 6 the Metafacet interface for the concordance of “hazard” is shown for the facet “author” sorted by frequency. Figure 6 shows a window of this data focusing on the nine most frequent attributes of this facet in the concordance list. The hover interaction is shown for “Thucydides”, who is the most frequent author of the keyword “hazard” in the GoK corpus, with a total of 94 concordance lines out of a list of 335.

Metafacet when used alone provides an interface to quickly explore keyword distribution across meta-data attributes. By interactively combining it with the concordance list and Mosaic we can navigate the corpus in a new way, viewing the concordance as attributed sets of collocations that can be interactively explored. In Figure 7 the stop-word Mosaic shown in Figure 2 is filtered to remove any concordance lines with the attribute “book” form the “format” facet. Books account for the majority of the concordance lines, and removing them from the concordance significantly changes the collocation structure of the Mosaic. During interactive filtering the current selection can be kept by pressing the “Update Bars” button, and this will refresh the Metafacet window with filtered concordance.

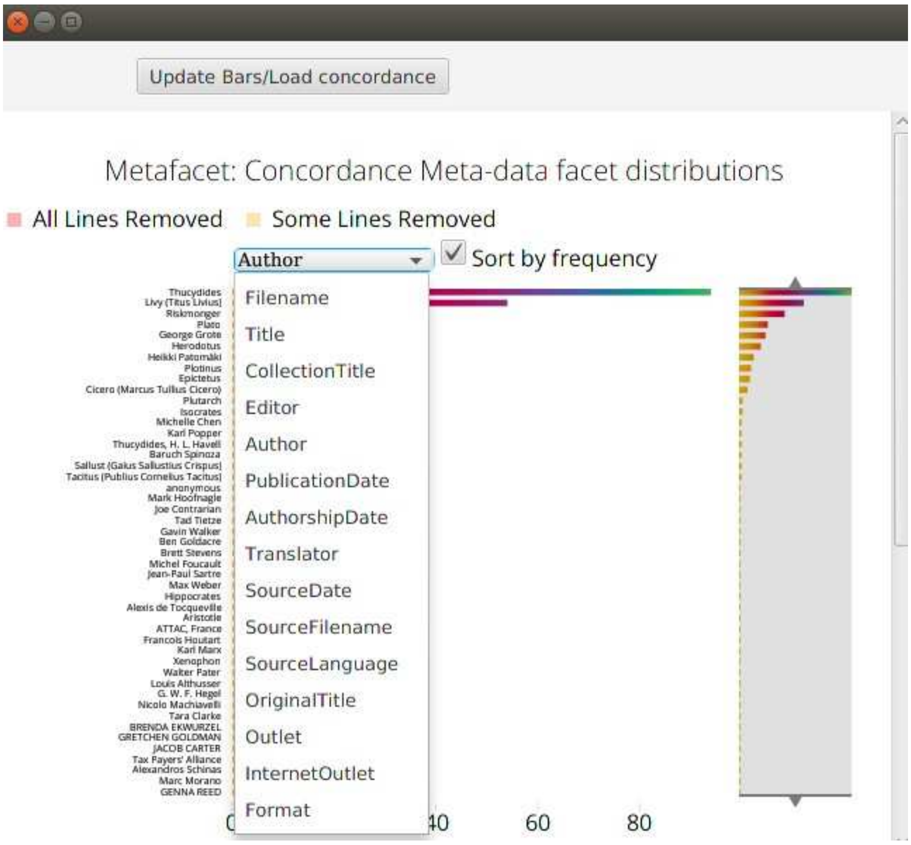


Figure 5: Metafacet interface showing all available meta-data facets. Fully zoomed out but obscured view of all authors in the concordance of "hazard".

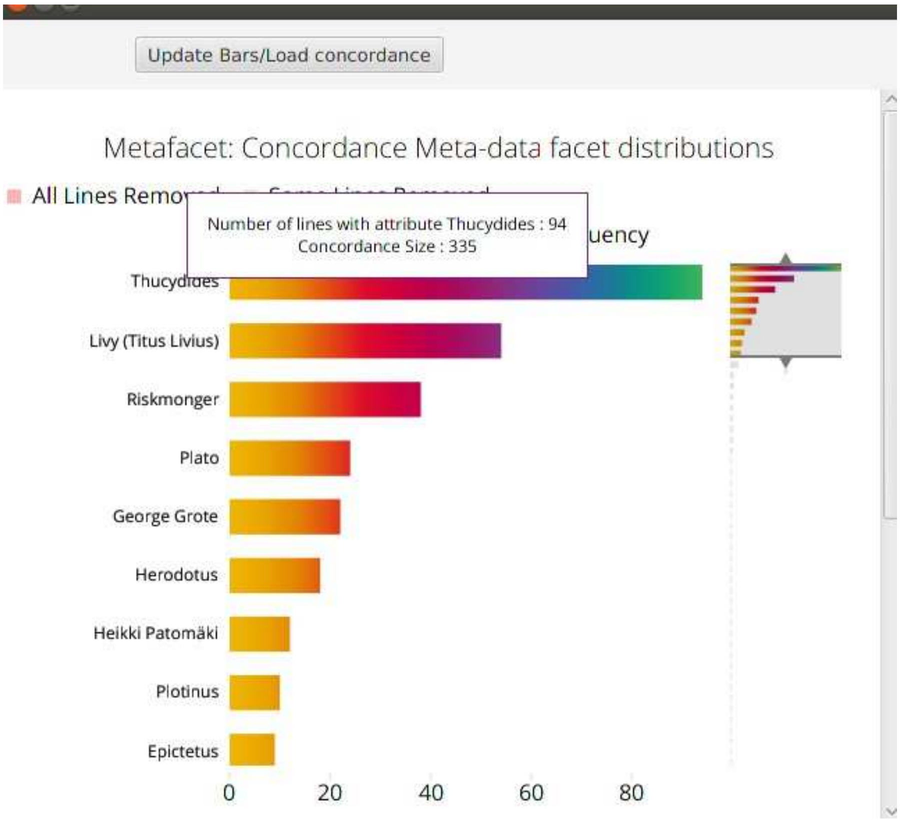
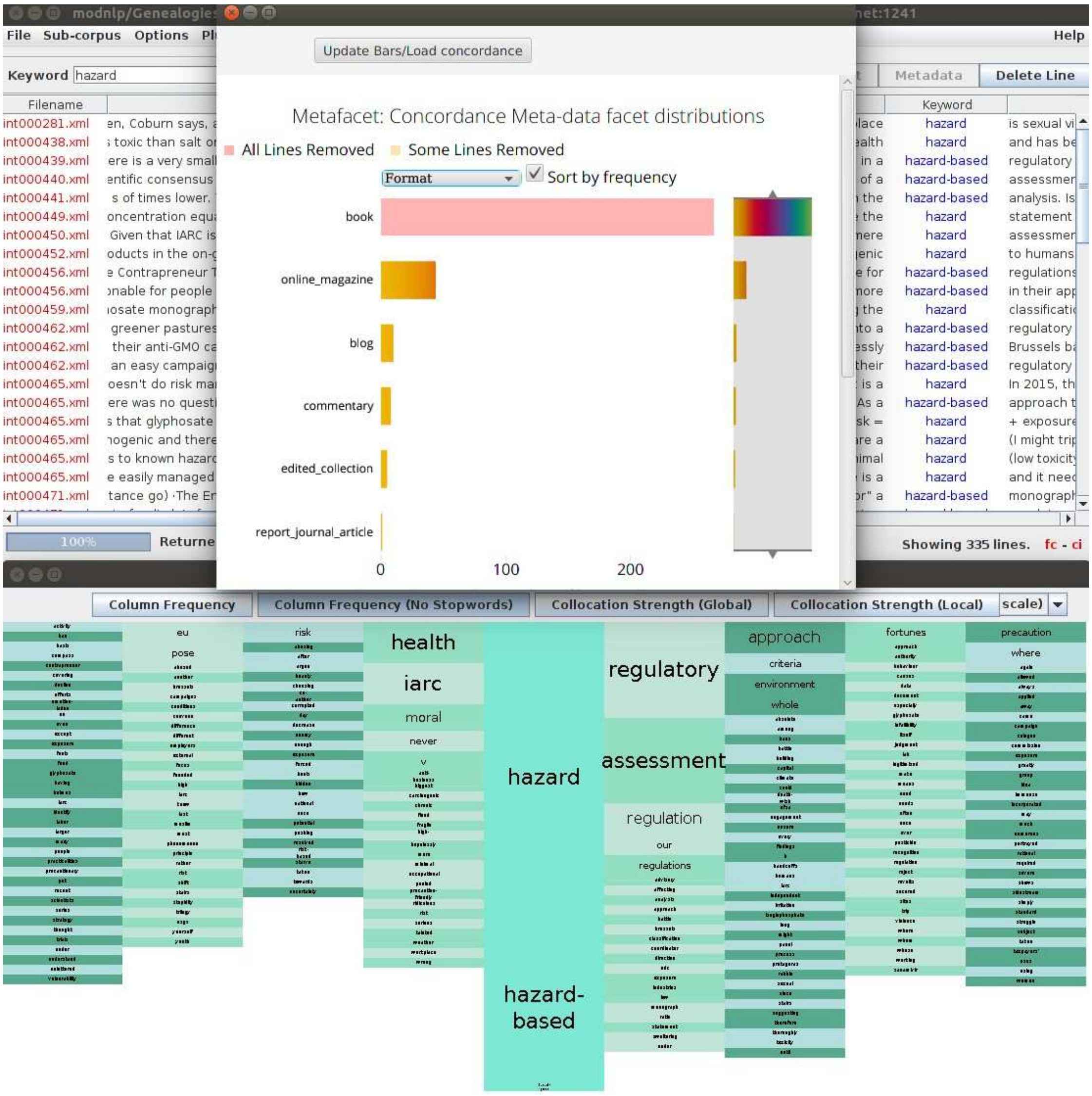

Number of lines with attribute Thucydides : 94
Concordance Size : 335

Figure 6: Metafacet zoomed to most frequent authors. Hover interaction displaying attribute name, associated concordance lines and total concordance lines for "hazard".



modnlp/Genealogies … net:1241

File Sub-corpus Options Pl… Help

Keyword hazard

Filename	
int000281.xml	en, Coburn says, a
int000438.xml	s toxic than salt or
int000439.xml	ere is a very small
int000440.xml	entific consensus
int000441.xml	s of times lower.
int000449.xml	oncentration equa
int000450.xml	Given that IARC is
int000452.xml	oducts in the on-g
int000456.xml	e Contrapreneur T
int000456.xml	onable for people
int000459.xml	iosate monograph
int000462.xml	greener pastures
int000462.xml	their anti-GMO ca
int000462.xml	an easy campaign
int000465.xml	oesn't do risk mai
int000465.xml	ere was no questi
int000465.xml	s that glyphosate
int000465.xml	nogenic and there
int000465.xml	s to known hazard
int000465.xml	e easily managed
int000471.xml	tance go) ·The Er

100% Returne…

	Keyword	
lace	hazard	is sexual vi
ealth	hazard	and has be
in a	hazard-based	regulatory
of a	hazard-based	assessmen
the	hazard-based	analysis. Is
the	hazard	statement
nere	hazard	assessmen
jenic	hazard	to humans
e for	hazard-based	regulations
nore	hazard-based	in their app
the	hazard	classificati
ito a	hazard-based	regulatory
essly	hazard-based	Brussels ba
their	hazard-based	regulatory
is a	hazard	In 2015, th
As a	hazard-based	approach t
sk =	hazard	+ exposure
re a	hazard	(I might trip
imal	hazard	(low toxicity
is a	hazard	and it need
or" a	hazard-based	monograph

Showing 335 lines. fc - ci

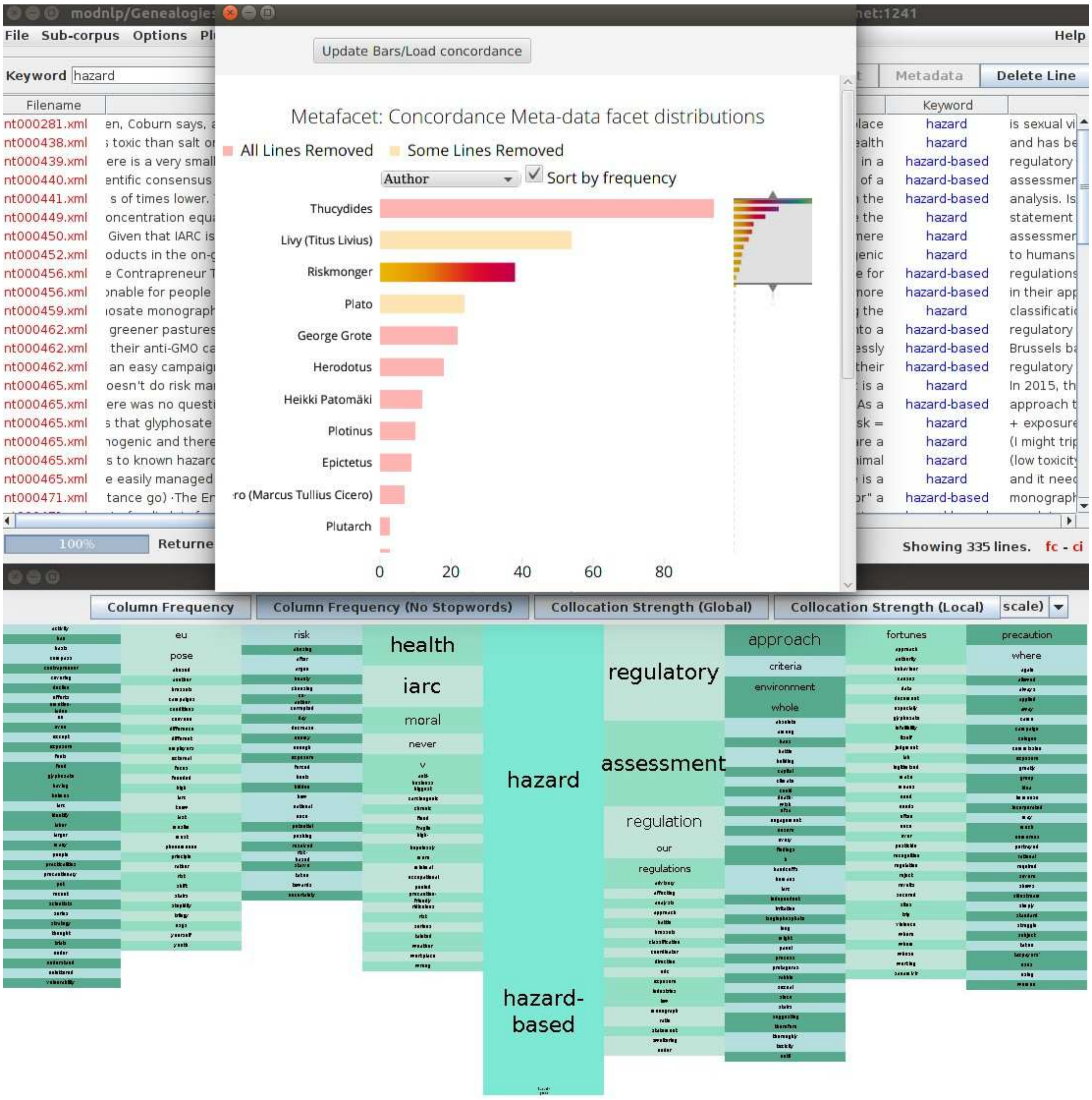
Metafacet: Concordance Meta-data facet distributions

■ All Lines Removed ■ Some Lines Removed

Format ☑ Sort by frequency

Figure 7: Left click interaction filtering out any lines form the concordance associated with the attribute Format= “book”. Both Concordance Mosaic and List respect the click interaction.

Left clicking a bar removes an attribute from the concordance list, right clicking removes everything but the clicked attribute. Once an attribute or multiple attributes have been selected it is possible to switch to another facet to explore further. In Figure 8 the facet “author” is displayed after books have been removed. We can see from the red bars that the most frequent author was only found in books. The second and fourth most frequent authors are coloured yellow, this indicates that some of the lines associated with these authors have been removed but others have not. To view how much these yellow bars have been reduced the “Update Bars” button must be pressed to generate a new Metafacet for the filtered concordance. It is possible on this author facet window to add attributes back into the list by clicking on the red or yellow bars, and this would generate a filtered list where all books except those of the selected authors have been removed.



Filename	
nt000281.xml	en, Coburn says,
nt000438.xml	; toxic than salt or
nt000439.xml	ere is a very small
nt000440.xml	entific consensus
nt000441.xml	s of times lower.
nt000449.xml	oncentration equ
nt000450.xml	Given that IARC is
nt000452.xml	oducts in the on-g
nt000456.xml	e Contrapreneur T
nt000456.xml	onable for people
nt000459.xml	iosate monograph
nt000462.xml	greener pastures
nt000462.xml	their anti-GMO ca
nt000462.xml	an easy campaig
nt000465.xml	oesn't do risk ma
nt000465.xml	ere was no questi
nt000465.xml	s that glyphosate
nt000465.xml	nogenic and there
nt000465.xml	s to known hazarc
nt000465.xml	e easily managed
nt000471.xml	tance go) ·The Er

	Keyword	
lace	hazard	is sexual vi
ealth	hazard	and has be
in a	hazard-based	regulatory
of a	hazard-based	assessmen
the	hazard-based	analysis. Is
the	hazard	statement
nere	hazard	assessmen
jenic	hazard	to humans
e for	hazard-based	regulations
nore	hazard-based	in their app
the	hazard	classificati
nto a	hazard-based	regulatory
essly	hazard-based	Brussels ba
their	hazard-based	regulatory
is a	hazard	In 2015, th
As a	hazard-based	approach t
sk =	hazard	+ exposure
re a	hazard	(I might trip
imal	hazard	(low toxicity
is a	hazard	and it need
or" a	hazard-based	monograph

Figure 8: Viewing frequent authors after filtering out attribute Format = "book". Partially removed attributes coloured yellow, fully removed attributes coloured red.

The combination of facets and attributes which can generate a single filtered list is limited only by the attribute crossover of the concordance lines. Finally the only author not colouring a block red or yellow in the nine most frequent is "Riskmonger", who does not have any concordance lines associated with the attribute "book". We stop the analysis here, but further exploration could be done to investigate the concordance lists and Mosaics for facets such as authorship/source dates and outlets. We would find that "Riskmonger" is a modern internet author who is responsible for the collocation patterns of "hazard" + "regulatory" and "assessment" at position K+1.

2.3 ComFre

The ComFre visualization is a corpus comparison tool where frequency lists can be compared visually in a statistically valid manner. The functionality of the tool has been detailed elsewhere (Sheehan et al., 2018), it has since been modified to operate as a plugin for Modnlp and is briefly presented here.

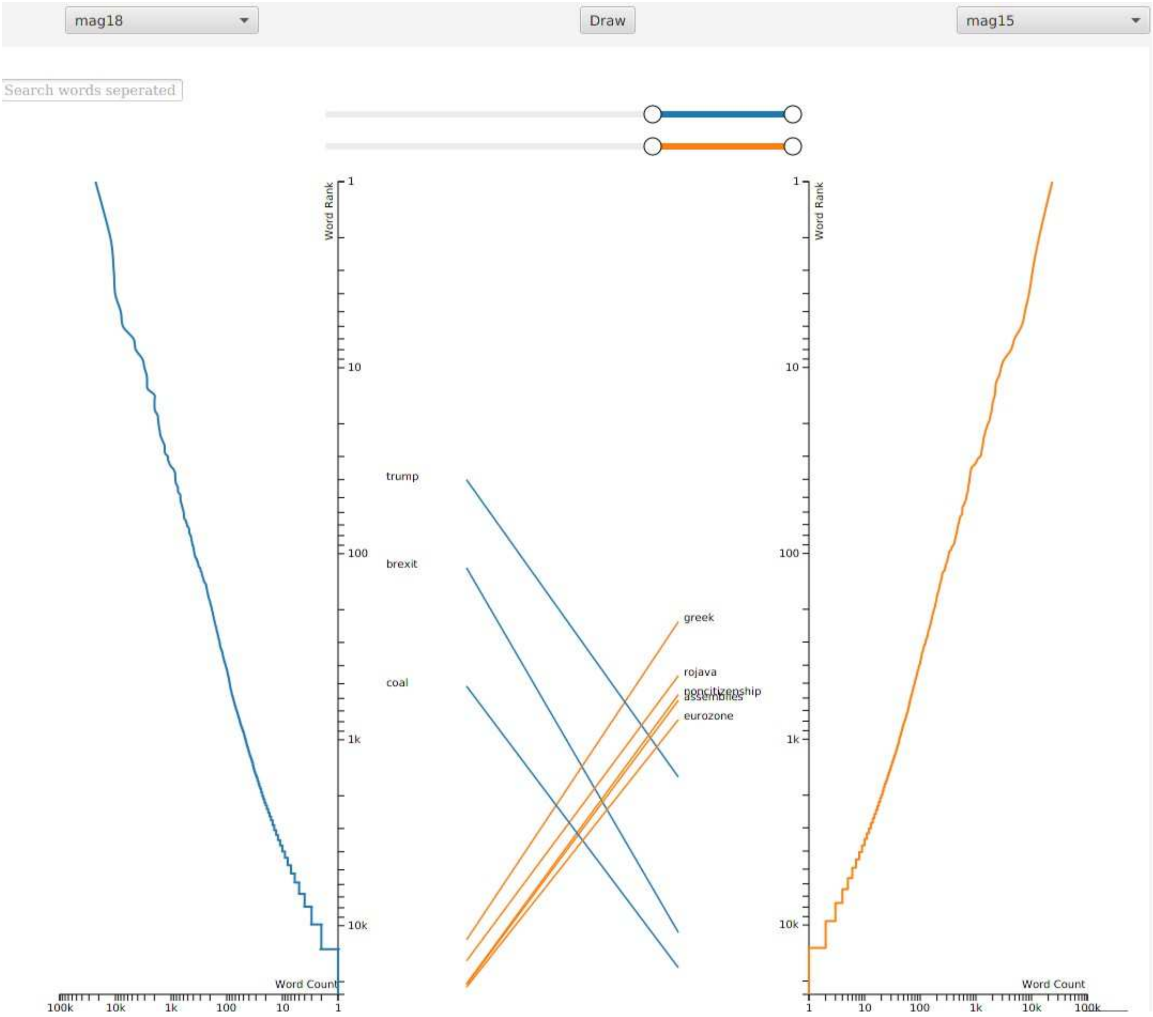


Figure 9: ComFre visualization comparing the words with the largest change in distribution rank between magazine articles from the GoK corpus authored in 2018 and 2015.

The Modnlp software has a sub-corpus selection interface which can be used to save the named sub-corpora for later reuse. ComFre makes these named subcorpora available for comparison in dropdown lists. In Figure 9 “mag18” and “mag15” are selected for comparison, these subcorpora are magazine articles from the GoK corpus which were authored in 2018 and 2015, respectively.

In ComFre both axis are log scaled, which should yield a linear frequency diagram if the word frequencies follow a Zipfian distribution. Scaling both ranked lists to the same height and comparing a word’s position in the distributions lets us compare subcorpora of vastly different size.

In Figure 9 the majority of the words have been filtered out to reveal the words with the greatest frequency changes between the two corpora. We can see that “Trump””, “Brexit”” and “coal”” were used much more often in the 2018 corpus, while words such as “Greek” and “eurozone” had much higher usage in 2015.

3. Domain characterization summary

This section explores the domain of corpus linguistics to identify problems and methods which will benefit from visualization. Visualizations which try to address the needs of corpus linguists are much more likely to be effective if those needs are well understood. The inclusion of domain experts in this visualization design stage is very beneficial, however just talking to users is typically not sufficient to achieve a full and accurate domain characterization. Expert users are extremely important when defining the high level goals and tasks of the domain and with ranking the importance of tasks. The characterization can be made more detailed by using methods such as examination of domain literature, contextual studies (Sedlmair et al., 2012) and needs assessments (Marai, 2018).

By performing a domain characterization, as outlined in the nested model (Munzner, 2009), the methodologies used to achieve the identified goals can be systematically investigated. The aim is to extract the low level tasks which are performed in the process of working towards the higher level goals. This analysis can be arranged as a hierarchy of goals, tasks and low level actions. The hierarchy can then be used to gain insight into the challenges faced by corpus linguists and how they have been previously addressed.

At its core domain characterization for visualization design is about identifying real problems which are relevant to the domain under investigation. This process is fluid and iterative, a level of domain understanding must be reached before work can begin on a visualization, but the design process should be reviewed as opportunities to refine the problems and domain characterization emerge.

The analysis presented here is not a full detailing of our characterization efforts. Rather, it is a presentation of some of the clearer insights and how they relate to the design choices which can be observed in the created visualization tools.

3.1 Literature-based domain analysis

Consultation and collaboration with the language scholars of the GoK project who interrogate corpora as an essential part of their analytical work lead to the natural discussion of visual tools to support analysis.

These collaborations revealed how integral the KWIC-based concordance display is to the work of the text analyst. These visual representations provide an essential view of the context in which the keyword occurs. However, examining the relative frequencies of the words which surround the keyword is also a commonly performed task using these tools, for which it would appear these tools are not well suited. In practice, the analyst usually complements the textual information provided by the KWIC display with lists of words sorted by frequency of occurrence in the sub-corpus under examination, as well as other statistics. Different processes and sub-tasks mediate the analysis as a whole.

To study this type of concordance analysis in a practical context we turned to a reference work entitled *Reading Concordances: An Introduction* (Sinclair, 2003). This book is intended as a tutorial on how to look for certain linguistic properties of a keyword (such as word sense, phrasal usage, part of speech and many others) using a KWIC concordance list. The reader is invited to perform eighteen tasks which introduce the key practical actions and usage of linguistic knowledge required to make decisions about the properties of a word or collocation. For each of these tasks we performed a hierarchical task analysis (Annett, 2003) by combining or splitting the steps into a series of actions and sub-actions.

Each of the eighteen tasks was analysed and tagged to assist with the classifying and counting of the actions and sub-actions. Before explaining the exact meaning of the tags, an example of the tagging procedure for task 4 is given. This tagging procedure can allow a visualisation researcher with limited knowledge in the problem domain to extract meaningful actions.

Task 4 is concerned with identifying literal and metaphorical usage phrases. The preamble to the task provides some linguistic insight explaining that "some idiomatic phrases in English are recognizable because they contain a word which is not found anywhere else, like *at loggerheads*". They may also be recognizable because the literal meaning is absurd. But others are more subtle and don't have the aforementioned identifying marks. As an example the phrase *he got cold feet* seems to be a literal way of saying that his feet are cold. How do we as readers know when it means he is cowardly? The task studies the example of the phrase "free hand". A concordance of 30 lines is provided and a set of twelve directions in how to analyses the concordance are given to the reader. An answer key is also provided which expands on the analysis and the insights that can be gained.

The first direction tells the reader to look at the position directly to the left of the phrases which have been sorted alphabetically "and list them in order of frequency. Can you associate any of the SINGLETONS with any of those that recur?" (Sinclair, 2003: 21) We tag this action with the *frequency* tag, *word position* tag, *group* tag and *expert decision* tag. The key gives a breakdown of the words at the position and notes that "her, your" are in the same word class as "his" and that "completely. fairly, totally" are in the same word class as "relatively".

Step two asks the reader to

> "Look again at the five lines where N—1 is an adverb of degree. What is the word at N—2? Then consider the two lines where N—1 is one. What is the word at N—2? Can you associate these seven lines with the two big groups of a and his . . . ?"

The positional notation N—2 means the set of words two positions to the left of the keyword. The same tags are applied to this action as word position, exact frequency counts and linguist knowledge are used. The answer key states

> "Where N - 1 is an adverb of degree, N—2 is a; so these five lines join the group of the indefinite article. Where N—1 is the word one, in no. 25 N - 2 is her and so this line joins those with possessive adjectives. The other one, no. 24, has only at N - 2 , which is unlike all the other lines in this sample, so we will fit it in later on."

Step three starts by explicating that in the previous step 28 of the 30 lines were extracted and divided into two groups based on "choice of determiner in front of the noun hand" the reader is then told "here the difference is not just the type of determiner; consider the meaning of free hand in the two types of line and comment on the distinction in meaning." This task is tagged with *Similar Meaning*, *expert decision* and *read context.* For this examples the meanings of the keyword must be analysed by reading the contexts and using linguist knowledge to compare the meanings The answer key explains that when a possessive adjective is the determiner the word "free" means "available" and the word "hand" is a part of the human body. When the determiner is a the phrase "a free hand" it means "an unrestricted opportunity".

Skipping forward to step seven the reader is narrowing in on the linguistic patterns which are used to determine literal or metaphorical usage of the phrase "free hand". The reader is asked to group concordance lines according to whether the verb is active or passive and to examine if this accounts for the use of the word "given" exclusively before "a free hand". Tags *group, read context and expert decision* all apply. Step 8 then combines all of the previous analysis to describe an algorithm for determining metaphorical or figurative usage of the phrase "free hand". Many of the lines which have been discarded as not matching any patterns are not included in the construction of the algorithm.

Condition 1 of the algorithm is that there is a form of the word "give" or a word with similar meaning to the left of the phrase. If not is there an occurrence of the verb "have" or "get", or one with a similar meaning and use?

Condition 2 is that the indefinite article precedes the core phrase, either directly or with only an adverb of degree in between.

If both conditions hold the phrase "free hand" means "to be set a task without restrictions on resources or methods to accomplish it".

Steps nine to twelve examine all that had not previously examined in the concordance. The word frequencies and patterns to the right of the keyword are analysed and used to help account for the lines which could not be explained by the left context analysis.

This example should help clarify how the tags were assigned to the individual steps of the tasks. There was a significant amount of variation across the tasks, but the core actions could be described with a relatively small set of tags.

The actions and sub-actions generalize the descriptive analysis steps into operations which are common to many of the tasks. Taking an overview of our classifications of these actions we created the hierarchy shown in Figure 10.

At the first level of the hierarchy, the primary actions (second level) are split into quantitative and qualitative groups. Qualitative actions are classified on the criteria that a decision, in which it would be possible for experts to disagree, needs to be made to complete the action. These experts could be human users or algorithmic classification processes. Quantitative actions may form a part of a qualitative action, for example, frequent patterns must be identified before they can be classified as phrasal or non-phrasal usage (Sinclair, 1991).

The quantitative actions are those in which the steps involved in the action can be clearly stated, and, given the classifications have already been made, the results will be the same when performed by a reliable analyst. For example, for a concordance word frequencies at a specific word position can be accurately and repeatably determined. The quantitative actions often make use of the results of a qualitative action, such as estimating the frequency of words to the left of a meaning group where the group has to first be identified by expert decision.

The second level of the hierarchy contains the primary actions. These are the actions which most often describe the spirit of the instructions given in the eighteen tasks. Deeper into the hierarchy the sub-actions required to perform these primary actions are presented.

At the third level of the hierarchy the *area of analysis* is displayed, this is the level at which we perform the primary action. Looking first at the quantitative actions, we found that in three of the primary actions (filter, frequency and estimate frequency) a word's position relative to the keyword is the area at which the actions are applied. A fourth quantitative action, frequent patterns, has an area of analysis, estimate frequency, which is one of the other primary actions. This means the action is performed on a collection of results from estimate frequency actions i.e. the analysis is performed on frequency estimations across word positions. It is worth noting that in four of the five quantitative tasks identified the word position or multiple word positions is the area at which the action is performed. The final action identified, *significant collocates*, uses the results of statistical analysis of the keyword and its context from the corpus under investigation. This analysis is usually undertaken as a separate piece of analysis, which has its results reported as a list of frequent collocations with a keyword.

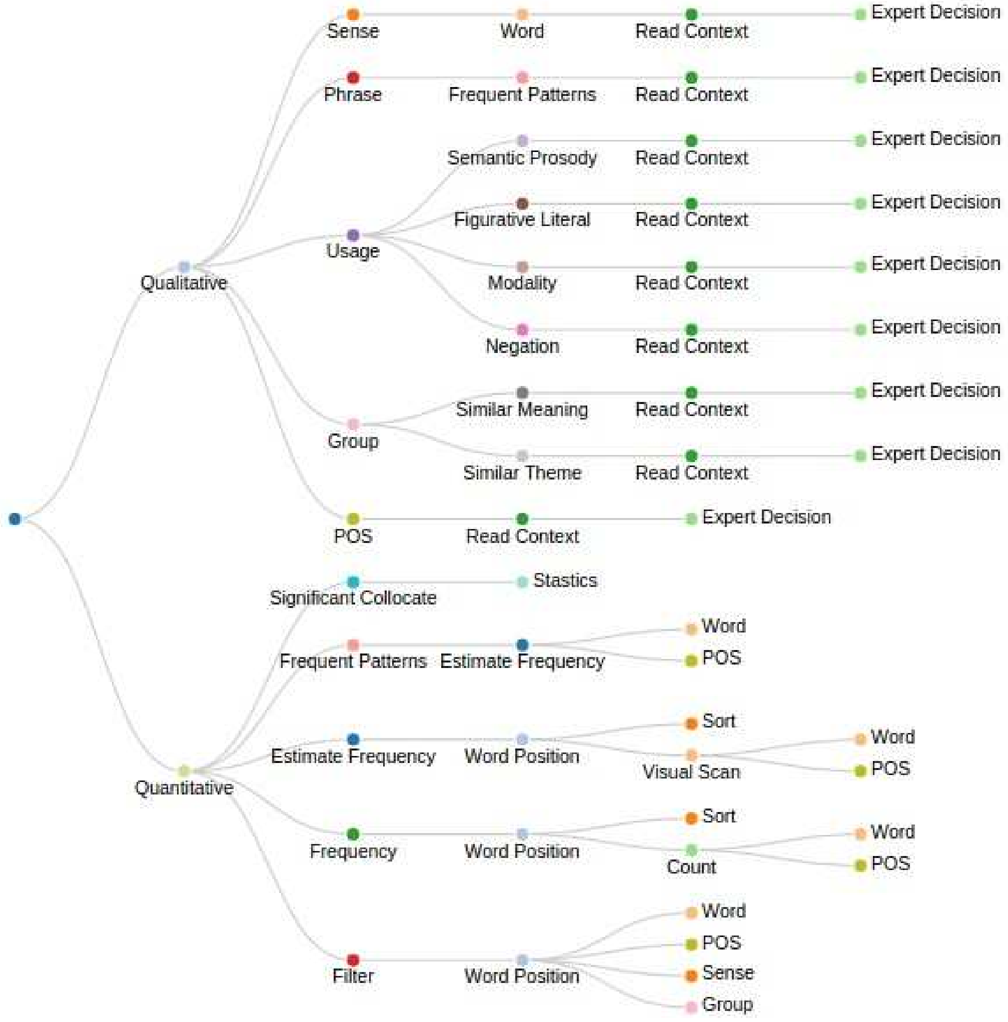


Figure 10: Hierarchical visualization of concordance-based corpus analysis actions.

Turning to the qualitative actions and, again, looking at the area of analysis at level three, we see that the analysis always occurs at the sentence level, which is implied by the read context action. This is in contrast with quantitative actions where positions are the most common area of analysis, and for qualitative actions it appears the horizontal structure of the KWIC list is emphasized while the qualitative actions make better use of the vertical alignment. Each of the actions requires an expert (or algorithm) who evaluates the context of individual occurrences of the keyword and makes a classification decision based on the semantic and syntactic content of the concordance line. This *Expert Decision* can often be the result of a combination of reading the individual contexts (the linear structure of the text) and performing some of the quantitative actions (positional statistics of the text). In essence, the *Expert Decision* action encapsulates the process of using the information extracted by other primary actions to answer questions about the keyword using linguistic knowledge.

Tag	No. of tasks in which an action appears	Total action appearances
expert decision	18	60
estimate frequency	16	34
read context	16	31
frequent patterns	15	21
frequency	14	18
word position	13	24
POS: Part of speech	11	23
filter	11	18
sense	10	19
group	7	9
significant collocate	5	7
usage	5	6
phrase	5	6

Table 1: Action counts from task analysis. Total numbers of actions found in the 18 tasks and numbers of the 18 tasks which feature the action.

While most of the tags represent actions, a few additional tags were chosen to help clarify and add information about the tasks and sub tasks. The tags *word*, *semantic prosody*, *Similar Meaning* and others are not themselves actions, but are useful in clarifying the objective or operation of the sub-actions. The part of speech (*POS*) tag is both a primary action tag and a clarifying tag. The POS primary action is to determine the part of speech of a word occurrence. The POS clarifying tag represents the use of part of speech information in another action. The purely clarifying tags are omitted from the analysis of tag frequency.

We recorded the distribution of the tags according to the number of tasks in which it appeared and the total number of actions which received the tag, as shown in Table 1. At a high level, this table tells us that both qualitative actions enabled by reading concordance lines and quantitative actions which require positional statistics are necessary for the style of concordance analysis outlined by Sinclair (2003).

3.1.1 Influence on visualization design

The structure that the task analysis and tag weightings add to the descriptive methodological steps was very useful for the early visualization design. The initial prototype of the Concordance Mosaic followed directly from this analysis. By focusing on frequent yet difficult aspects of the methodology we were able to create an interface which was likely to be of interest to corpus linguists. This gave us the opportunity to engage with domain experts in the iterative development of tools and methodology starting with a useful prototype.

3.2 Methodological descriptions for GoK case studies

During the development of the visualization tools many interactions with GoK researchers occurred in situations such as progress meetings, design reviews and informal meetings. One set of interactions which made significant contributions to identifying relevant domain problems is presented here.

This takes the form of an initial presentation and follow up observation session with one GoK researcher. In the initial meeting a simplified methodology for a case study was described and a visualization which could be useful was suggested. The follow up observation session took place a number of months later after the Mosaic interface had been improved and made available to the researchers.

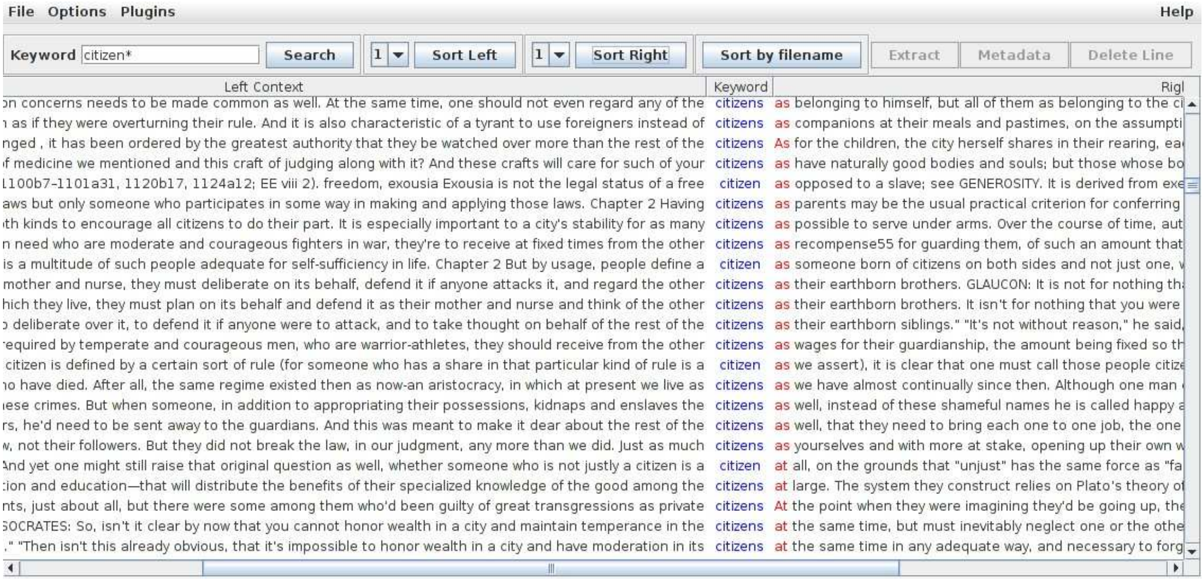
3.2.1 Methodology presentation

In the methodology discussion meeting a brief presentation outlining an example methodology and its challenges was given by a member of the GoK project to help with the initial definition of visualization goals for the project. The methodology was explained in the form of a case study. The case study made use of the portion of GoK English corpus which was available at the time. The task was defined as comparing the patterning around the keyword "*citizen**" . The * represents a regular expression search for continuations of the word citizen such as citizens and citizenship. The patterns identified were compared across two large sub-corpora.

- **Sub-corpus 1** A sub-corpus of modern English translations from Classical Greek (1850 onwards);
- **Sub-corpus 2** A sub-corpus of translated and non-translated texts written by contemporary authors, published between 1992 and the present day.

The method itself consisted of two techniques. The goal of the first technique is the identification of explicit definitions of "citizenship" contained within each sub-corpus. To find these definitions the researcher wants to compile a list of frequently used verbs and prepositions at position "keyword+1". To achieve this the GoK corpus browser is used. Sub-corpus 1 was selected using the sub-corpus selection tool, the regular expression "citizen*" was searched and the concordance was sorted at position "keyword+1". The researcher then spends time scrolling through the concordance and compiling a list of relevant frequent words at the position of interest, Figure 11 shows the concordance window sorted and scrolled to the preposition *as*. With this list in hand more accurate searches can be run such as:

- citizenship+“(is/as/was/defined/conceived/are/equals /considered/appears/means)”
- citizenship+“(has/should/must/will/may)”
- citizen+“(is/as)”
- citizens+“(are/as)”



Left Context	Keyword	Rigl
on concerns needs to be made common as well. At the same time, one should not even regard any of the	citizens	as belonging to himself, but all of them as belonging to the ci
1 as if they were overturning their rule. And it is also characteristic of a tyrant to use foreigners instead of	citizens	as companions at their meals and pastimes, on the assumpti
nged , it has been ordered by the greatest authority that they be watched over more than the rest of the	citizens	As for the children, the city herself shares in their rearing, ea
f medicine we mentioned and this craft of judging along with it? And these crafts will care for such of your	citizens	as have naturally good bodies and souls; but those whose bo
1100b7-1101a31, 1120b17, 1124a12; EE viii 2). freedom, exousia Exousia is not the legal status of a free	citizen	as opposed to a slave; see GENEROSITY. It is derived from exe
aws but only someone who participates in some way in making and applying those laws. Chapter 2 Having	citizens	as parents may be the usual practical criterion for conferring
th kinds to encourage all citizens to do their part. It is especially important to a city's stability for as many	citizens	as possible to serve under arms. Over the course of time, aut
n need who are moderate and courageous fighters in war, they're to receive at fixed times from the other	citizens	as recompense55 for guarding them, of such an amount that
is a multitude of such people adequate for self-sufficiency in life. Chapter 2 But by usage, people define a	citizen	as someone born of citizens on both sides and not just one, w
mother and nurse, they must deliberate on its behalf, defend it if anyone attacks it, and regard the other	citizens	as their earthborn brothers. GLAUCON: It is not for nothing th
hich they live, they must plan on its behalf and defend it as their mother and nurse and think of the other	citizens	as their earthborn brothers. It isn't for nothing that you were
o deliberate over it, to defend it if anyone were to attack, and to take thought on behalf of the rest of the	citizens	as their earthborn siblings." "It's not without reason," he said,
equired by temperate and courageous men, who are warrior-athletes, they should receive from the other	citizens	as wages for their guardianship, the amount being fixed so th
citizen is defined by a certain sort of rule (for someone who has a share in that particular kind of rule is a	citizen	as we assert), it is clear that one must call those people citize
10 have died. After all, the same regime existed then as now-an aristocracy, in which at present we live as	citizens	as we have almost continually since then. Although one man
iese crimes. But when someone, in addition to appropriating their possessions, kidnaps and enslaves the	citizens	as well, instead of these shameful names he is called happy a
rs, he'd need to be sent away to the guardians. And this was meant to make it dear about the rest of the	citizens	as well, that they need to bring each one to one job, the one
w, not their followers. But they did not break the law, in our judgment, any more than we did. Just as much	citizens	as yourselves and with more at stake, opening up their own w
And yet one might still raise that original question as well, whether someone who is not justly a citizen is a	citizen	at all, on the grounds that "unjust" has the same force as "fa
tion and education—that will distribute the benefits of their specialized knowledge of the good among the	citizens	at large. The system they construct relies on Plato's theory of
nts, just about all, but there were some among them who'd been guilty of great transgressions as private	citizens	At the point when they were imagining they'd be going up, the
SOCRATES: So, isn't it clear by now that you cannot honor wealth in a city and maintain temperance in the	citizens	at the same time, but must inevitably neglect one or the othe
." "Then isn't this already obvious, that it's impossible to honor wealth in a city and have moderation in its	citizens	at the same time in any adequate way, and necessary to forg

Figure 11: Visualization proposed by GoK researcher

By reading the concordance lines generated by these new searches definitions can be extracted. Some examples of the definitions found are:

- Citizenship is a status bestowed on those who are full members of a community.
- As well as enjoying rights, citizens are required to undertake responsibilities such as paying taxes, and jury or military service.
- Citizenship should be based purely on residency
- US citizenship has represented a safe haven from oppressive regimes around the world

The second technique is the observation of patterns in the kinds of adjectives used to modify “citizenship”, as well as constructions such as “citizens+of+*”. The researcher explained that this technique is more difficult and time consuming using a concordance browser. To quote the researcher.

> “Specifically, it is difficult to get a quick overview of such patterns using the concordancer given that the number of lines returned for my searches is quite large:

> e.g. 4420 hits for "citizen*" in my sub-corpus of translations from Classical Greek."

The researcher had some experience with linguistic visualization having used early versions of Mosaic and in the past had used word clouds, such as Wordle (Viegas et al., 2009), to present research results. There are some challenges to overcome to use word clouds for the methodology. The first which the researcher noticed is that stop-words dominate the frequency distributions of the word positions, so some technique has to be applied to get meaningful results. The suggested technique was to use a stop-word list to filter the visualization. The concordance would need to be processed to extract the words at particular positions for visualization, since the concordance is structured as a list of aligned text extracts. The result of the researchers reasoning was an interface for displaying positional word clouds with the option to exclude stop-words. The presentation included a mock-up of what a visualization to solve this problem would look like, as shown in Figure 12. The mock-up displays a word cloud for either a full concordance or a chosen word position, and has the option to remove stop-words. Looking at the mock-up in Figure 12 the words modifying citizen are presented in a manner that emphasizes frequency and provides an overview on a single screen of a position relative to the keyword.

At the end of the presentation the idea and its feasibility were discussed and some questions were asked to clarify the methodology. The notes taken were later discussed with the researcher and the following questions and answers were prepared.

- What is the domain in which the case study is situated?

  > "Translation and Reception studies. How have we received classic Greek texts? How has translation shaped this reception? The role of translation is often overlooked."

- Is this methodology (excluding the proposed visualization) typical of the field?

  > "Translation Studies as a discipline tends to encourage close qualitative analysis of a small selection of examples chosen from specific texts to illustrate a particular argument.
  >
  > Corpus analysis enables the translation scholar to identify and investigate with significantly greater ease differences between and patterns within translations, taking into account the full length of each work as a complete text.
  >
  > Corpus analysis has been extensively used in translation studies before (e.g. within the TEC project and many others) but the field has tended to focus mainly on more micro-level linguistic concerns, rather than the socio-political implications of translators' word-choices etc."

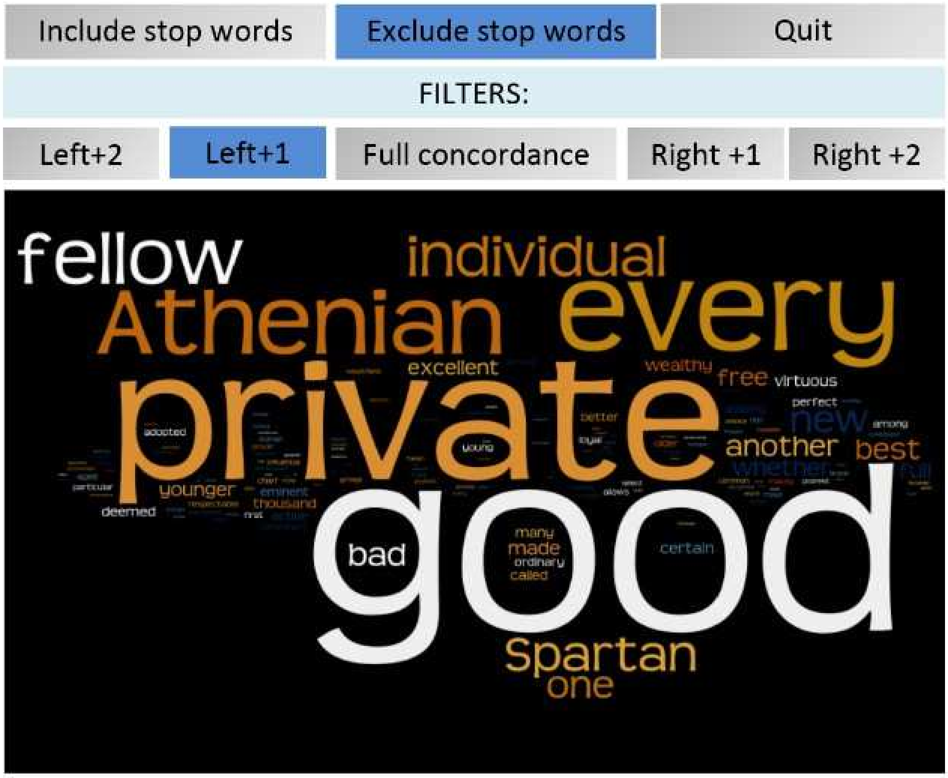


Figure 12: Visualization proposed by GoK researcher.

- How did the idea for this example arise?

> “GoK seeks to understand the constellation of concepts related to the body politic across time and space. Citizenship is a lexical item in that constellation. Comparing meaning, frequency and usage of related terms is an exploratory process used to discover obvious patterns.”

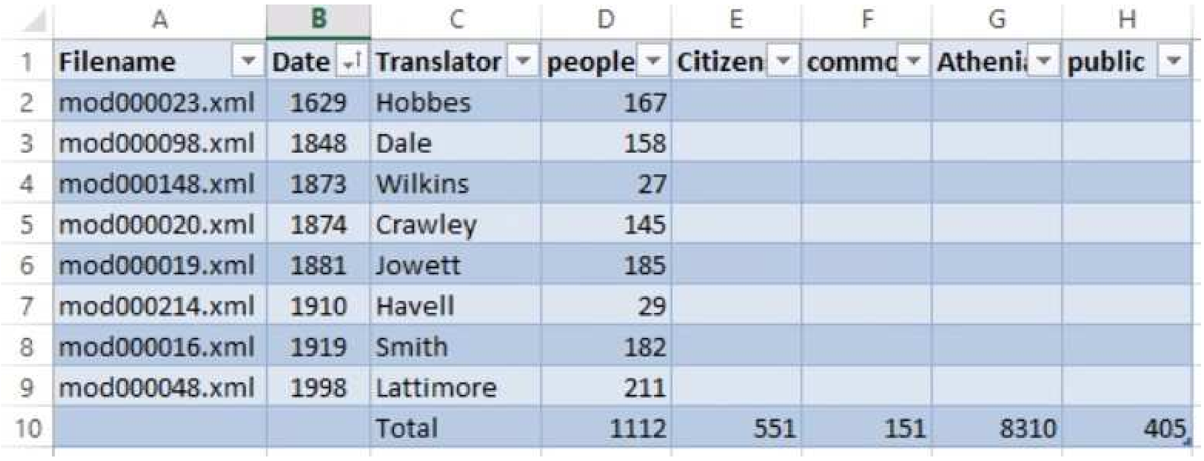
3.2.2 Methodology presentation: Design influence

The presentation helped confirm that the tasks and actions identified in the task analysis were relevant to at least one linguistics researcher. The early design of Mosaic did not take into account the need for removal of stop-words to make the Mosaic more usable. Th researcher identified this flaw but did not notice the equivalence between a mosaic column and a word cloud. By removing the stop stop-words from the Mosaic you present the same information as a positional word cloud with a greater visual emphasis on word position and frequency. This was a very beneficial meeting, and led to the addition of this “No Stop-word” view of Concordance Mosaic.

3.2.3 Methodology observation: Case study of “the people”

After a significant amount of time follow-up observation sessions were organized to gain further insight into the methodologies of the researcher who gave the presentation. This took place after the development and release of the mature Concordance Mosaic, but prior to the development of Metafacet.

Prior to the observation session a spreadsheet was created with the headings filenames, date, translator, people, citizens, commons, Athenians, public. The meta-data information related to filename, date and translator were added to the table. The remaining headings are keywords which will be investigated as part of this study. The spreadsheet used in the study can be seen in Figure 13. Partitioning the frequencies by date, file or translator is equivalent for this sub-corpus, as each file has a unique author and date.



	A	B	C	D	E	F	G	H
1	Filename	Date	Translator	people	Citizen	commc	Atheni	public
2	mod000023.xml	1629	Hobbes	167				
3	mod000098.xml	1848	Dale	158				
4	mod000148.xml	1873	Wilkins	27				
5	mod000020.xml	1874	Crawley	145				
6	mod000019.xml	1881	Jowett	185				
7	mod000214.xml	1910	Havell	29				
8	mod000016.xml	1919	Smith	182				
9	mod000048.xml	1998	Lattimore	211				
10			Total	1112	551	151	8310	405

Figure 13: The spreadsheet which was used in the study of “the people” in translations of “Thucydides” from the GoK corpus.

The first steps of the study focused on the keyword frequencies in the entire sub-corpus.

- The sub-corpus of “Thucydides” was selected.
- The keyword “people” was searched and the total frequency in the corpus was recorded
- Regulator expressions for the other “citizens?”, “commons?”, “Athenians” and “public” were searched and the total frequency in the sub-corpus was recorded.

The researcher commented, after the keyword frequencies had been recorded, that the keyword “Athenians” is much more frequent than other keywords. This is unexpected and will need to be investigated.

The next step was to gather the keyword frequencies for individual files.

- Make a sub-corpus selection for each individual file. Record in the spreadsheet the number of lines returned for the keyword “people”.

The analysis now turns from keyword frequency to the identification of collocation patterns. Mosaic was used extensively to identify collocation patterns and frequency of occurrence. The steps observed were:

- Make a sub-corpus selection for the first file.
- Perform a search for the first keyword “people” in the concordance browser.
- Open the Mosaic visualization and remove stop-words.
- Examine word frequencies.
- Open a document for taking notes and record in it the most frequent collocations directly to the left of the keyword. The words “common and “Athenian” were recorded.
- Return to the sorted concordance list and check if any continuations (such as “Athenians”) are present.
- Record the counts for the frequent collocated words. (common 8, Athenian 6).
- Open the frequency mosaic with stop-words included.
- Record in notes “lots of hits for the+people (i.e. unmodified)”
- Similar analysis for second file.
- Frequent collocates directly to left of “people” (common 34, Athenian 5).
- Record “A few more different adjectives modifying this noun:entire, experienced, free, dynamic, adventurous.”
- Similarly for the third file the noted collocates were (Athenian 13, whole 13, common 5).

The recording was ended and the researcher explained how the analysis would progress. The collocation pattern method is repeated and would continue in the same manner for each file and keyword. The next stage of the analysis would be to analyse the frequency patterns using the table. Possibly making bar charts in a spreadsheet application. Temporal patterns are expected. Identified patterns will be investigated using qualitative analysis, which involves reading the concordance lines related to the identified patterns. Understanding the meaning of the concept of “the people” at different times is the goal.

This analysis is performed in the context of the knowledge the researcher has about the corpus and texts. She states that it is interesting that there are

> "No translations 1919-1998, during period of huge cultural change in Britain. Possible reasons for this include Suffrage, war or technological revolution. The researcher explained that information about the authors and texts will influence the analysis. Some examples of information which is relevant are "the political leanings of the translators which is established relevant knowledge" and "certain texts are partial translations, abridged versions etc."

Any differences identified, temporal or otherwise, must take into account translator style, politics and more.

Some questions were asked the researcher to elicit more information about the methodology

- How did you come up with this methodology?

  > "Playing around with the corpus tools, generating concordances for interesting keywords, trying to find patterns in the data."

- How did you choose the keywords?

  > "Obvious keywords associated with the concept of "the people". The idea for the study emerged through reading the literature on citizenship."

- Would this methodology be useful for other researchers in the field?

  > "Other scholars using the GoK software to investigate the role of translation in the evolution of political and scientific discourse use similar methods. Other projects developing other corpora may also adopt some aspects of the methodology."

- What are barriers to the adoption of your methodology?

  > "Not sure. Perhaps better documentation of the corpus software, detailing what it can and can't do, with lots of example analysis. The publication of case-studies by members of the team will also help demonstrate the potential of the tools."

- Mosaic was used in this analysis, is this typical when you investigate collocation patterns?

  > "Yes. Mosaic will be very useful for this case-study and any investigation of collocations, because it tells you in very quick and transparent way which

> are the most common collocates in each word position for a given keyword."

- You did not make use of collocation strength in your analysis, do you intend to?

> "No. The collocation strength Mosaic is not immediately clear, and so (to be brutally honest) would tend to slow down analysis rather than speed it up."

- Have you used this methodology for other studies?

> "The collocation pattern aspect of this study is unique in my work. I have in previous studies studied keyword frequency in larger sub-corpora where there are multiple files for each author and date. I can show you an example for the concept of "Statesman"."

3.2.4 Methodology observation: Case Study of "Statesmanship"

An unpublished paper on a case study of the concept of "Statesmanship" was supplied by the researcher and the major conclusions and analysis were described.

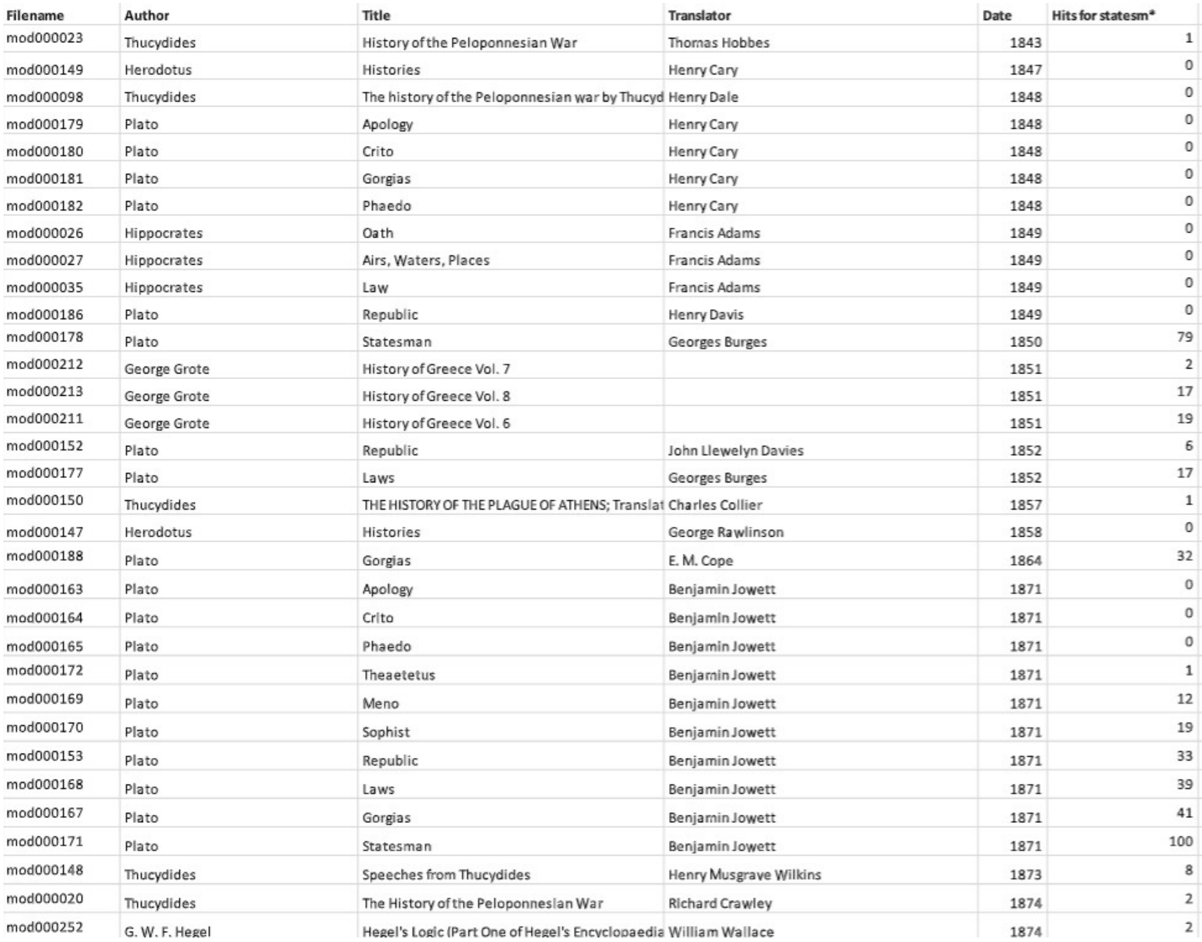
In the GoK corpus the term "statesman" was found to exist "almost exclusively (90%) in translations from Classical Greek". This pattern was not observed for other similar keywords such as "governor", "leader", "ruler" and "citizen", which are more evenly distributed across all language pairs. The analysis which arrived at this conclusion was a simple keyword frequency comparison across the translation facets of the corpus. This involved selecting each sub-corpus individually and recording the number of concordance lines for the keywords in each sub-corpus.

The frequency of the keyword "statesman" in the sub-corpus of Classical Greek translations was analysed. A spreadsheet with an entry for each of the 261 files in the sub-corpus was created and meta-data (the author, the title, the translator and the date) was entered for each file. This was done manually and was time consuming. The researcher explained that in this form "the information could easily be (re)sorted according to each of these meta-data facets and patterns more easily identified". The number of concordance lines for each file was found by selecting a sub-corpus of a single file and searching for "statesman". Performing this action for each of the 261 files was also time consuming. A sample of the completed spreadsheet can be seen in Figure 14.

By examining the spreadsheet and generating bar charts, such as Figure 15, the faceted distributions of "Statesman" can be understood. "statesman" seemed to be "bursty", to use the author's term, and to exhibit a temporal pattern.

> The frequency of "statesman" in these corpora suggest most recent translations (1950-2012) of ancient Greek texts use "statesman" much less frequently. This is surprising because the corpus contains several recent re-translations

(published within the last seventy years) of classical texts such as Aristotle's Politics or Plato's Dialogues which in earlier English-language interpretations included the keyword "statesman" very prominently.



Filename	Author	Title	Translator	Date	Hits for statesm*
mod000023	Thucydides	History of the Peloponnesian War	Thomas Hobbes	1843	1
mod000149	Herodotus	Histories	Henry Cary	1847	0
mod000098	Thucydides	The history of the Peloponnesian war by Thucyd	Henry Dale	1848	0
mod000179	Plato	Apology	Henry Cary	1848	0
mod000180	Plato	Crito	Henry Cary	1848	0
mod000181	Plato	Gorgias	Henry Cary	1848	0
mod000182	Plato	Phaedo	Henry Cary	1848	0
mod000026	Hippocrates	Oath	Francis Adams	1849	0
mod000027	Hippocrates	Airs, Waters, Places	Francis Adams	1849	0
mod000035	Hippocrates	Law	Francis Adams	1849	0
mod000186	Plato	Republic	Henry Davis	1849	0
mod000178	Plato	Statesman	Georges Burges	1850	79
mod000212	George Grote	History of Greece Vol. 7		1851	2
mod000213	George Grote	History of Greece Vol. 8		1851	17
mod000211	George Grote	History of Greece Vol. 6		1851	19
mod000152	Plato	Republic	John Llewelyn Davies	1852	6
mod000177	Plato	Laws	Georges Burges	1852	17
mod000150	Thucydides	THE HISTORY OF THE PLAGUE OF ATHENS; Translat	Charles Collier	1857	1
mod000147	Herodotus	Histories	George Rawlinson	1858	0
mod000188	Plato	Gorgias	E. M. Cope	1864	32
mod000163	Plato	Apology	Benjamin Jowett	1871	0
mod000164	Plato	Crito	Benjamin Jowett	1871	0
mod000165	Plato	Phaedo	Benjamin Jowett	1871	0
mod000172	Plato	Theaetetus	Benjamin Jowett	1871	1
mod000169	Plato	Meno	Benjamin Jowett	1871	12
mod000170	Plato	Sophist	Benjamin Jowett	1871	19
mod000153	Plato	Republic	Benjamin Jowett	1871	33
mod000168	Plato	Laws	Benjamin Jowett	1871	39
mod000167	Plato	Gorgias	Benjamin Jowett	1871	41
mod000171	Plato	Statesman	Benjamin Jowett	1871	100
mod000148	Thucydides	Speeches from Thucydides	Henry Musgrave Wilkins	1873	8
mod000020	Thucydides	The History of the Peloponnesian War	Richard Crawley	1874	2
mod000252	G. W. F. Hegel	Hegel's Logic (Part One of Hegel's Encyclopaedia	William Wallace	1874	2

Figure 14: A sample from the spreadsheet used in the study of "statesman" in translations of Classical Greek from the GoK corpus. The full spreadsheet contains 261 lines of analysis.

Some clarifying questions were asked and answered:

- You mentioned the process of completing the spreadsheet was time consuming, how long did it take?

> "Probably around 5-6 hours because of the amount of manual processing required. It would take a lot longer if I were to investigate more than one keyword."

- Where did the idea for this study and methodology come from?

> "This was exploratory. I was not trying to establish anything in particular, only to understand whether the term "statesman" was used, how frequently (in comparison with other semantically related terms), and if any obvious patterns could be found from these initial quantitative analyses.

> The terms “statesman” and “citizenship”, which I have investigated previously, are very closely related concepts, especially in classical Greek thought.”

- Were the visualization tools used in this case study?

> “My focus on the use of a single keyword (“statesman”) and alternative word choices did not require and collocation pattern analysis. This is more typical of translation studies research. The corpus tools lend themselves particularly well to the analysis of collocations (this is one of their clear advantages), and this is why I want to push my research in this direction with my next case study.”

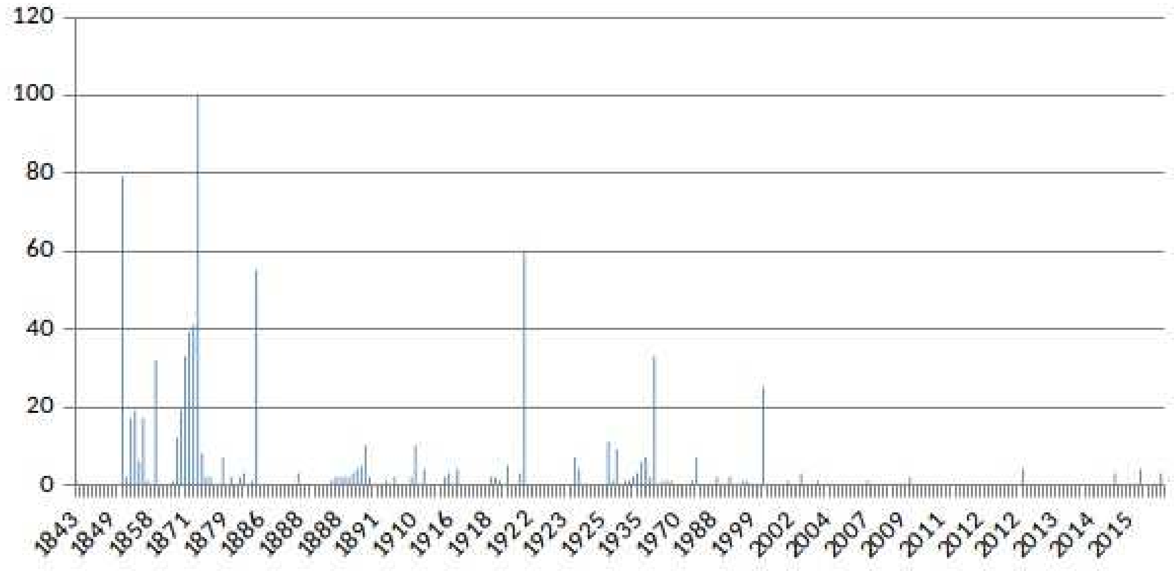


Figure 15: Bar chart examining temporal spread in translations of ancient Greek.

- Are there any areas of your methodology where you current or new visualization tools could be beneficial?

> “Constructing the spreadsheets is time consuming. A tool which can help identify patterns in the dispersion of a concept according to different meta-data facets would be extremely helpful, at least for the kinds of research I intend to carry out as part of this project.”

3.2.5 Case study observation: Influence on visualization design

The most significant outcome of the two case studies was the emergence of the obvious need for a method to support the analysis of concordance lists through the lens of metadata. This observation session led to further discussion and needs assessment for a meta-data analysis tool which eventually became Metafacet.

Another problem identified was that in the version of Mosaic available to the researchers at that time only a single collocation statistic was available, and it was based on Mutual Information. The researcher did not know exactly what the scaling scheme for the collocation strength of Mosaic View was, and so could not accurately interpret or use it for publication. This led to the creation of optional scaling schemes based on well-known collocation metrics. More collocation measures are still being added to the tool.

4. Discussion and conclusions

We have presented three visualization techniques for corpus analysis. We hope that they can be adopted where appropriate by lexicographers and the wider corpus linguistic community. In addition, discussion of the tools and techniques by the community is welcomed.

We would be glad to hear any ideas, comments or criticisms of our ideas, understanding and designs. We believe the problems the tools address are general enough to have wide applicability in corpus linguistics, but we do not doubt that specific domains, such as lexicography, will have nuanced requirements that may need specialized interactions or entire redesigns to make them useful enough to be widely adopted.

The domain characterization detailed here can be another point of discussion, perhaps leading to more specialized future work on specific domain problems. We believe it is extremely important to provide a rationale for design decisions and to engage with domain experts when designing or modifying a tool or technique. Future work in this area will take the form of modifications which are identified during further domain exploration, and new visualization techniques where entire new problem areas are uncovered.

5. Acknowledgements

This paper is supported by European Union's Horizon 2020 research and innovation programme under grant agreement No 825153, project EMBEDDIA (Cross-Lingual Embeddings for Less-Represented Languages in European News Media). The results of this paper reflect only the author's view and the Commission is not responsible for any use that may be made of the information it contains. Pierre Albert has been funded though the INCA project. We thank the INCA project members in Ireland for granting us access to the trainee data.

6. References

Annett, J. (2003). Hierarchical task analysis. *Handbook of cognitive task design*, 2, pp. 17–35.

Bonelli, E. T. (2010). Theoretical overview of the evolution of corpus linguistics. *The*

Routledge handbook of corpus linguistics, p. 14.

Doherty, G., Karamanis, N. & Luz, S. (2012). Collaboration in Translation: The Impact of Increased Reach on Cross-organisational Work. *Computer Supported Cooperative Work (CSCW)*, 21(6), pp. 525–554.

Karamanis, N., Luz, S. & Doherty, G. (2011). Translation practice in the workplace: contextual analysis and implications for machine translation. *Machine Translation*, 25(1), pp. 35–52. URL http://dx.doi.org/10.1007/s10590-011-9093-x.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P. & Suchomel, V. (2014). The Sketch Engine: ten years on. *Lexicography*, 1(1), pp. 7–36. URL http://dx.doi.org/10.1007/s40607-014-0009-9.

Luhn, H. & Division, I.B.M.C.A.S.D. (1959). *Keyword-in-context Index for Technical Literature (KWIC Index)*. ASDD Report. International Business Machines Corporation, Advanced Systems Division. URL http://books.google.ie/books?id=Dk7pAAAAMAAJ.

Luz, S. (2000). A Software Toolkit for Sharing and Accessing Corpora Over the Internet. In M. Gavrilidou, G. Carayannis, S. Markantonatou, S. Piperidis & G. Stainhauer (eds.) *Proceedings of the Second International Conference on Language Resources and Evaluation: LREC-2000*. pp. 1749–1754.

Luz, S. (2011). Web-based corpus software. In A. Kruger, K. Wallmach & J. Munday (eds.) *Corpus-based Translation Studies – Research and Applications*, chapter 5. Continuum, pp. 124–149.

Luz, S. & Masoodian, M. (2007). Visualisation of Parallel Data Streams with Temporal Mosaics. In E. Banissi et al. (eds.) *Procs. of the 11th International Conference on Information Visualisation*. Zurich: IEEE Computer Society, pp. 197–202.

Luz, S. & Sheehan, S. (2014). A Graph Based Abstraction of Textual Concordances and Two Renderings for their Interactive Visualisation. In *Proceedings of the International Working Conference on Advanced Visual Interfaces*, AVI '14. New York, NY, USA: ACM, pp. 293–296.

Manning, C. D. & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. Cambridge, Massachusetts: The MIT Press.

Marai, G. E. (2018). Activity-Centered Domain Characterization for Problem-Driven Scientific Visualization. *IEEE Transactions on Visualization and Computer Graphics*, 24(1), pp. 913–922.

Munzner, T. (2009). A Nested Model for Visualization Design and Validation. *IEEE Transactions on Visualization and Computer Graphics*, 15(6), pp. 921–928.

Rockwell, G. (2003). What is Text Analysis, Really? *Literary and Linguistic Computing*, 18(2), pp. 209–219.

Scott, M. (2010). What can corpus software do. *The Routledge handbook of corpus linguistics*, pp. 136–151.

Sedlmair, M., Meyer, M. & Munzner, T. (2012). Design Study Methodology: Reflections from the Trenches and the Stacks. *IEEE Transactions on Visualization and Computer Graphics*, 18(12), pp. 2431–2440.

Sheehan, S., Masoodian, M. & Luz, S. (2018). COMFRE: A Visualization for Comparing Word Frequencies in Linguistic Tasks. In *Proceedings of the 2018 International Conference on Advanced Visual Interfaces*, AVI '18. New York, NY, USA: ACM, pp. 42:1–42:5. URL http://doi.acm.org/10.1145/3206505.3206547.

Shneiderman, B. (1996). The eyes have it: a task by data type taxonomy for information visualizations. In *Proceedings the IEEE Symposium on Visual Languages*. pp. 336–343.

Sinclair, J. (1991). *Corpus, concordance, collocation*. Describing English language. Oxford University Press.

Sinclair, J. (2003). *Reading Concordances: An Introduction*. Longman Publishing Group. URL http://books.google.ie/books?id=Ms9nQgAACAAJ.

Summers, D. (1996). Corpus lexicography–the importance of representativeness in relation to frequency. *Longman Language Review*, 3, pp. 6–9.

Viegas, F. B., Wattenberg, M. & Feinberg, J. (2009). Participatory Visualization with Wordle. *IEEE Transactions on Visualization and Computer Graphics*, 15(6), pp. 1137–1144.